How to restore nature after dams come down p. 254

A brain-inspired computer chip pp. 263 & 329

A very distant fast radio burst p. 294

Science

$15
20 OCTOBER 2023
science.org

AAAS

COPIOUS CICADAS

Mass emergence alters food webs pp. 268 & 320

EDITORIAL

Safeguard the world's worst pathogens

Piers Millett
is the executive director of the International Biosecurity and Biosafety Initiative for Science, Washington, DC, USA. piers@ibbis.bio

More people in more places are researching the world's most dangerous pathogens. This work helps prepare against future pandemics, but it's not without danger. Pathogens could escape from research facilities, so it is vital for countries to assess potential hazards and have procedures in place to manage the risks. This doesn't happen enough.

Earlier this year, the Global Biolabs initiative released a report tracking the growth of maximum containment labs. These facilities provide extremely high levels of protection when there are very high risks to lab personnel, the wider community, or the environment. The report reveals that in 2000, there were only 13 such labs in the world. The number of labs in operation, under construction, or planned has risen steadily: to 59 labs in 23 countries in 2021, and 69 labs in 27 countries in 2023. This boom in labs has not been sufficiently accompanied by strengthened safety and security.

> "Every country... needs a plan to prevent accidents and deliberate harm."

At the international level, there is guidance to help prevent harm from maximum containment labs, including guidance from the World Health Organization on lab biosafety (2020) and responsible conduct (2022), and from the World Organization for Animal Health on lab biosafety and biosecurity (2023) and responsible conduct (2019). Compliance, however, remains voluntary, and keeping pace with technical developments remains a challenge. A stronger evidence base for assessing risks and the efficacy of responses is needed. For example, the 2022 White House lab biorisk science & technology roadmap needs to move forward, in particular the recommendation to create a globally distributed research agenda.

Although maximum containment–based research is intended to provide benefits, it could be misapplied to do harm. The risks of such "dual use" research deserve more international attention. There has been a need for a global organization—such as the International Biosecurity and Biosafety Initiative for Science, whose primary focus is on the emerging risks from biotechnology—that can help to "de-risk" tool development and implementation, and help to bring together policy, academic, and public sectors. At the national level, every country with a maximum containment lab must have relevant rules and regulations in place. It should be clear what research can (and cannot) be done, who is responsible, and what standards are to be followed. Every country with such a lab needs a plan to prevent accidents and deliberate harm. Very few countries have dedicated biosecurity strategies, such as the National Biological Security Strategy released by the UK in June. Every country with a maximum containment lab should be cultivating a strong culture of responsibility, such as is found in the International Genetically Engineered Machine competition. Responsible oversight and strengthening cultures of responsibility should also be integrated into national bioeconomy strategies—which are much more widely adopted than their biosecurity counterparts.

Does the world really need so many of these labs? Equitable access will be vital to create a world where local people can solve local problems using the powers of bioscience and biotechnology. However, aside from the risks, maximum containment labs are costly and complex to run. Alternative strategies already exist for reducing the need for containment when researching dangerous pathogens. For example, sometimes parts from pathogens can be studied in less risky microbes, as is increasingly common in basic research and vaccine development for influenza viruses and coronaviruses.

Future capabilities may come online that further erode the need for these labs. Machine learning is allowing more work to be attempted on computers. This does not completely remove the need for labs, though it might reduce the number required to verify simulations with real experiments. This approach opens the door for more distributed access to research. Policy-makers are concerned that these tools could be misused to cause deliberate harm. The potential for artificial intelligence (AI) to aid the development of bioweapons will be discussed at the AI Safety Summit. A more permanent forum is needed to understand the biosecurity implications of the overlap of AI and bioscience. Better and more widely adopted tools to screen for potential hazards are needed in gene synthesis companies, biofoundries, and cloud labs—which will help turn simulations into biology.

As more people in more places research dangerous pathogens, it is vital that they manage the risks. Balancing the need for critical research with measures to ensure the work is done safely, securely, and responsibly is a task for everyone.

–Piers Millett

PHOTO: UNIVERSITY OF OXFORD

10.1126/science.adl3600

> “Many researchers ... used such collections to advance deeply flawed scientific agendas rooted in white supremacy.”
>
> **American Museum of Natural History President Sean Decatur,** in *The New York Times,* about the museum’s plan to reconsider its historic collection of human remains, which includes bones of Indigenous and enslaved people.

IN BRIEF

Edited by **Jeffrey Brainard**

Technicians use a glovebox as they work to extract asteroid material collected in a sampling container.

PLANETARY SCIENCE

Asteroid samples rich in water, organics

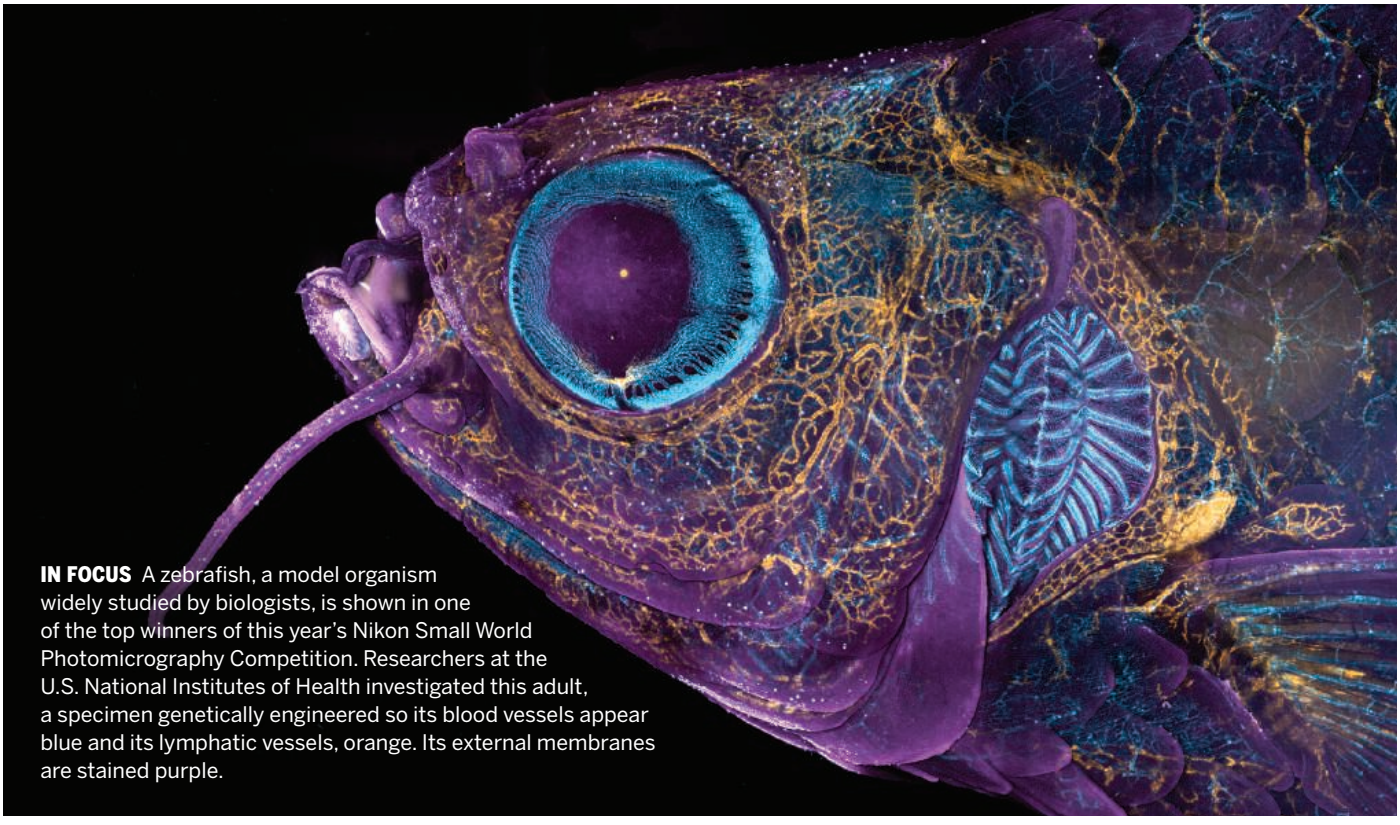
The first look at material from the asteroid Bennu, delivered to Earth last month by NASA’s OSIRIS-REx spacecraft, has revealed high levels of water and organic molecules. Further analysis could yield clues about how asteroid impacts on Earth billions of years ago may have helped life get a start. Technicians at Johnson Space Center have not yet opened the main sample container. But initial tests of dust and grit that leaked into a surrounding compartment show hints of carbonate, which may have precipitated out of flowing water on Bennu’s parent body during the infancy of the Solar System, NASA said last week. After the container is opened and the roughly 250 grams of asteroid material it holds are cataloged, one-quarter will be distributed to scientists worldwide and the rest preserved for future generations.

China tightens ethics reviews

RESEARCH INTEGRITY | China last week further toughened government regulation of biomedical research by requiring ethics reviews of studies done in the private sector. The new measures also clarify review procedures for work involving animals, artificial intelligence, medicine, and genetic data. Previous rules, unveiled in March, covered only research involving human subjects conducted at public institutions. The new rules were prompted, in part, by the controversial genetic engineering of three babies in 2018 by biophysicist Jiankui He, who now works at Wuchang University of Technology and a private nonprofit in Beijing. But, “China still has quite some distance to go to enforce these new measures,” says sociologist Joy Zhang of the University of Kent. And a preprint posted last week by a group at the Second Nanning People’s Hospital suggests past regulations have had limited impact. It reported that, in a survey, more than 53% of 6200 medical residents at 17 hospitals in southwest China admitted committing at least one form of research misconduct.

Wellcome Trust gets new head

LEADERSHIP | John-Arne Røttingen has been named the new head of the Wellcome Trust, one of the world’s largest nongovernmental funders of scientific research, the U.K. foundation announced last week. Røttingen, who holds M.D. and Ph.D. degrees, was the founding CEO of the Coalition for Epidemic Preparedness Innovations and has been ambassador for global health at the Ministry of Foreign Affairs in his native Norway since 2020. During the pandemic, he led the steering committee of the World Health Organization’s (WHO’s) Solidarity Trial, which evaluated drugs for treating SARS-CoV-2 infections. After he takes the reins at Wellcome in 2024, Røttingen will help oversee the spending of £16 billion (more than $19 billion) over the next decade to find science-based solutions to challenges such as mental health, infectious disease, and the health impacts of climate change, the foundation said. He will succeed Paul Schreier, who became interim CEO this year after Jeremy Farrar, Wellcome’s previous head, stepped down to become WHO’s chief scientist.



IN FOCUS A zebrafish, a model organism widely studied by biologists, is shown in one of the top winners of this year's Nikon Small World Photomicrography Competition. Researchers at the U.S. National Institutes of Health investigated this adult, a specimen genetically engineered so its blood vessels appear blue and its lymphatic vessels, orange. Its external membranes are stained purple.

PHOTO: DANIEL CASTRANOVA AND BRANT WEINSTEIN/NIKON SMALL WORLD PHOTOMICROGRAPHY COMPETITION

AI spots, reports supernova

ASTRONOMY | Researchers have applied an automated method to detect a new supernova without human intervention, which they say will reduce error and speed up discoveries. The first signs of the supernova, called SN 2023tyk, were recorded on 3 October by the Zwicky Transient Facility, an automated telescope network in California that surveys the sky and typically identifies about 50 bright candidate events a night. Astronomers often need days to manually inspect them and get follow-up observations of the most promising. Instead, a team at Northwestern University applied an artificial intelligence algorithm, called the Bright Transient Survey Bot, which had been trained on 1.4 million historical images. It requested automated follow-up observations for SN 2023tyk and by 7 October had reported the confirmed and classified supernova to the Transient Name Server, a clearinghouse for astronomers. Supernovae last for weeks, but researchers want to observe them early to obtain clues about processes such as how they forge heavy elements.

Scientists decry 'antiwoke' vow

POLICY | A pledge by the United Kingdom's Conservative government to "kick woke ideology out of science" has sparked controversy. In a 3 October speech, science minister Michelle Donelan vowed to stop "the steady creep of political correctness" in research. She claimed that scientists are being "told by university bureaucrats that they cannot ask legitimate research questions about biological sex" and said the government is launching a review to produce guidance in 2024 on sex and gender questions in studies. More than 2000 researchers have signed an open letter stating they "are appalled, saddened, and angry that the government claims to be speaking for science in this way." It accuses Donelan of trying to "stoke so-called culture wars and make the U.K. increasingly hostile towards people identifying as intersex, non-binary, and/or trans. … Far from depoliticizing science, this policy appears to be driven by ideology."

BY THE NUMBERS

10^{40}

Number of cells that have ever lived, according to one of the first studies to give an estimate. The total is 22 orders of magnitude higher than the number of sand grains on Earth. (*Current Biology*)

High court declines carbon case

CLIMATE POLICY | The U.S. Supreme Court last week refused to hear a challenge to the Biden administration's plan to assign a higher cost to the damage caused by carbon pollution. The plaintiffs—a coalition of Republican-led, fossil fuel-producing states—argued that the administration's plan to raise what's called the social cost of carbon would harm their economies. The metric attempts to capture the costs of carbon emissions, such as adverse health effects, that are not reflected in market prices; agencies use it to calculate the costs and benefits of regulations. The Biden administration previously raised it to $51 per ton and this year proposed a further increase to $190. Industry groups and others are expected to challenge specific agency rules that use the estimate.

Project collects African genomes

BIOMEDICINE | Researchers this week announced plans to create the first-ever database of genomes exclusively from people with African ancestry. They are among the most genetically diverse in the world but remain underrepresented in genetic databases, and by extension, genetic studies: Only 0.48% of participants in such research worldwide identify as African American or Afro-Caribbean and

a mere 0.18% as African. As a result, some treatments—for cancer, for example—may not work well in this group because they were tested primarily in people of European ancestry. The new effort, called Together for Changing Healthcare for People of African Ancestry through an InterNational Genomics & Equity, aims to recruit at least 500,000 volunteers. It is being led by Meharry Medical College, a historically Black college, and the pharmaceutical companies AstraZeneca, Novo Nordisk, Regeneron, and Roche. Each company is contributing $20 million. The group expects to form partnerships with other historically Black U.S. institutions and with universities in Africa. All partners will have exclusive access to the resulting data.

Ancient Egyptian snakes ID'd

ARCHAEOLOGY | Scientists have put names to some of the mysterious snake species that bit ancient Egyptians. Researchers have long disagreed about the identities of 34 snakes mentioned in a scroll, dated to the 6th century B.C.E. and housed in the Brooklyn Museum, that describes how to treat snakebites and is considered one of the first medical texts. A team at Bangor University analyzed 19 variables that describe the climate of known snake species' ranges and concluded that nine—including the highly venomous puff adder and black mamba—that no longer live in modern Egypt could have lived in the climate prevailing in ancient Egypt, they reported on 7 October in *Environmental Archaeology*.

Wildcats reintroduced in Scotland

CONSERVATION | The first deliberate reintroduction of a mammalian predator in the United Kingdom is showing signs of success as 19 Scottish wildcats appear to be adapting to the wild, the Royal Zoological Society of Scotland announced last week. It bred the critically endangered animals—a subpopulation of the Eurasian wildcat (*Felis silvestris*)—and released them this summer into Cairngorms National Park. Data from GPS transmitters and wildlife cameras indicate they are hunting.

Laser lines illuminate a scroll burned by the ancient eruption of Mount Vesuvius.

ARCHAEOLOGY

AI and x-rays decipher burned scroll

Researchers last week unveiled the first whole word to be identified from what remains of papyrus scrolls incinerated by the eruption of Mount Vesuvius in 79 C.E.: "porphyras," which means "purple" in ancient Greek. Unfurling the scrolls, recovered from a ruined library in the ancient town of Herculaneum, would almost certainly destroy them. But the researchers were able to read letters inside one of them by training an artificial intelligence algorithm to decipher the grainy, barely legible characters revealed by x-rays. By decoding that first word, Luke Farritor, a computer science student at the University of Nebraska, Lincoln, earned a $40,000 prize from a competition, the Vesuvius Challenge, which began in 2019. Still unclaimed is its $700,000 grand prize for the first to decipher four passages from inside one of the scrolls; the deadline looms at the end of this year.

THREE QS

Reform-minded editor may shake up psychology journal

After high-profile retractions and reports that many studies can't be replicated, behavioral science is under scrutiny. So proponents of reform welcomed the news this month that a leading proponent of rigor and open science will lead one of the field's most prominent journals, *Psychological Science,* in January 2024. The new editor-in-chief, University of Melbourne psychologist Simine Vazire, has stood up for change, resigning in 2018 from a U.S. National Academies of Sciences, Engineering, and Medicine panel on reproducibility over how much error in science should be considered acceptable. She spoke with *Science* about her plans for the journal. (A longer version of this interview is at https://scim.ag/VazireQA.)

Q: Why take these reins?
A: I've worked a lot from within the establishment while also being prepared to call out things that I see. I have to admit that I was surprised [at being chosen]. But it feels nice to see that even a journal that's quite mainstream is open to continuing to move in this direction.

Q: Why is this important for reform?
A: [*Psychological Science*] controls a lot of incentives. Getting a paper into a [prestigious] journal can affect your chances of getting a grant, a job, an award. So even small changes in the incentives and reward structure at journals have a really big impact.

Q: What do you intend to change?
A: I plan to move to transparent peer review—publishing the [anonymous] peer review history for all accepted manuscripts. Editors can help make verification of findings easier. We're planning to require that any published article has open data, materials, and code, with exceptions in cases where that's ethically challenging. Journals can encourage either internal or third-party checks, maybe random checks or end-of-year audits. Journals can do more to also create a culture where just because something passed peer review doesn't mean that's the end of scrutiny, and where we welcome and facilitate ongoing questioning.

SCIENCE.ORG/NEWS
Read more news from *Science* online.

IN DEPTH

In August, the Sun rose b[e]hind the Extremely Large Telescope, under construction in Chile.

PHOTO: EDUARDO GARCES/EUROPEAN SOUTHERN OBSERVATORY

ASTRONOMY

Giant European scope rises as U.S. rivals stumble

Past the halfway point, Extremely Large Telescope prepares to receive first mirrors

By **Daniel Clery**

A web of steel girders is rising from the flattened summit of Cerro Armazones, 3000 meters above sea level in Chile's Atacama Desert. The dome it will support will be vast—with a footprint as big as a soccer field and almost as tall as the Statue of Liberty—and unexpectedly nimble: It will smoothly rotate on rails as a giant telescope inside tracks stars through the night.

Everything about the aptly named Extremely Large Telescope (ELT) inspires awe. Its main reflector will be a bowl of silvered glass 39 meters across—the size of three IMAX screens—and the patchwork of 798 hexagonal mirror segments that comprise it must together form a perfect parabola down to a few tens of nanometers. The biggest optical telescope ever built, it will take pictures of Earth-like worlds around others stars and look for signs of life in their atmospheres.

A few months ago, the European Southern Observatory (ESO), which is building the $1.5 billion telescope, declared that construction had passed the halfway mark. Next month, the first batch of polished mirror segments should set off from France by ship in a temperature-controlled container. "It really is going to happen," says astronomer Richard Ellis of University College London.

As ELT marches toward completion in 2028 or 2029, U.S. astronomers can only look on in envy. A pair of similarly ambitious U.S.-led projects—the Giant Magellan Telescope (GMT) and the Thirty Meter Telescope (TMT)—once vied with ELT to be first on the sky. Although the projects are also polishing mirrors, they have not begun construction—and success is not assured. Unable to find enough funding from private and international sources, they are waiting for the National Science Foundation (NSF) to bail them out by paying at least 25% of their combined cost of about $5 billion. Earlier this month, NSF Director Sethuraman Panchanathan, speaking to a Senate committee, said the agency was looking to make a budget request for construction funding—but not until fiscal years 2025 or 2026. "There is some urgency," says John O'Meara, chief scientist of the W. M. Keck Observatory. "They need investment by NSF to move forward or the projects are in serious jeopardy."

ESO, with 16 member nations and a treaty that guarantees steady annual income, can pay for ELT without serious sacrifices, although a few years ago it had to shrink the mirror from an originally planned 42 meters. But for TMT and GMT, led respectively by the California Institute of Technology and the University of California and by the Carnegie Institution for Science, it's been a struggle. Fundraising has stalled even though the projects have attracted international partners in Australia, Brazil, South Korea, Japan, India, Canada, and China. Ellis, a former TMT board member, thinks it was unrealistic for private institutions to think they could manage multibillion-dollar projects. "This is well beyond any individual university or consortium," he says.

The two projects acknowledged as much in 2018. Burying the hatchet after years of competition, they joined forces to propose that NSF take a share in the two telescopes, giving publicly funded U.S. astronomers access. In 2021, that proposal, dubbed US-ELT, was named the highest priority in ground-based astronomy by the field's "decadal survey," a community exercise that is meant to guide NSF and Congress.

Both projects passed NSF's preliminary design review earlier this year. But that is just the first of many hurdles. Next week, an NSF panel will assess whether the projects are ready to move to final design and will make a recommendation to Panchanathan. Only after the projects are in the final design stage will NSF consider requesting construction funding from Congress.

Even for the U.S. government, the ask may be too big. Just one of the telescopes would be NSF's biggest ever outlay for a facility, and Ellis isn't sure NSF's astronomy division can convince others that contributing to both at a cost of well over $1 billion is justified.

Another problem is operating costs, which at NSF come out of the same account as grants to researchers. Funding one-quarter of US-ELT's day-to-day operations—likely to be more than $100 million annually—would create a painful squeeze for the astronomy division. Ellis thinks NSF's slowness to act is a sign of reluctance. "After 2 years there's nothing to show," he says. "It would be a tragedy for U.S. astronomy if they don't get built."

In the case of TMT, there is an added complication. TMT's chosen site, near the summit of Mauna Kea on Hawaii's Big Island, has been opposed by local people including Native Hawaiians who want no more telescopes built on what they consider a sacred place. Protesters blocked an attempt to begin construction in 2015. A second attempt in 2019 was also stymied after protesters built a camp along the summit access road.

Now, says TMT Executive Director Robert Kirshner, "We're doing a lot of community work in Hawaii. ... We have to find a solution that is agreeable to everyone." NSF, too, is legally required to assess the project's environmental, historical, and cultural impacts before endorsing it. Officials held public meetings on the Big Island in the summer of 2022, some marked by fractious debates. The agency is now evaluating what it learned. Linnea Avallone, NSF's Chief Officer for Research Facilities, says the process will take 2 years to complete.

TMT and GMT continue to work with what funding they have. Three of GMT's seven 8.4-meter mirrors are finished, three are awaiting polishing, and technicians at the University of Arizona last month loaded 20 tons of glass into a circular mold to cast the last one. The glass will be heated gradually to 1165°C while slowly spinning to give the mirror its parabolic shape.

At the GMT site at Las Campanas Observatory in Chile the seven mirrors will be combined to form a 25.4-meter main reflector. It will give GMT a wide field of view, perfect for surveys, says Chief Scientist Rebecca Bernstein. "Some kinds of science need to get large numbers of objects very efficiently," she says. The support structure for this mirror is fully designed, Bernstein says, and the team is preparing to begin construction.

Like ELT, TMT will consist of a mosaic of smaller mirrors, an approach pioneered on the existing 10-meter Keck telescopes. At the Richmond, California, plant of the company Coherent, an assembly line is polishing three 1.5-meter mirrors each month before trimming them into hexagons. While Coherent is making 230 mirror segments, other teams in India, China, and Japan will produce the remainder of TMT's required 492 (plus 80 spares). Kirshner says 97 are done so far. Canada is poised to build the telescope's enclosure and Japan has committed to building its support structure—once the site dispute has been settled.

German company Schott has cast glass "blanks" for the Extremely Large Telescope's 798 hexagonal mirror segments.

ELT has had hiccups of its own. The €400 million contract for its structure and enclosure—the biggest ESO has ever awarded—went to a consortium of two Italian companies, Astaldi and Cimolai. When Astaldi went bankrupt in 2018, Cimolai had to take over, which led to delays that ESO says it has recovered from. French company Safran-Reosc is producing up to four mirror segments a week at a dedicated factory in Poitier and is ramping up, says ELT program manager Roberto Tamai. Once the first 45 have passed inspection, they'll be packaged up and dispatched to Chile. Tamai says: "Now you can touch with your hand things that were only on paper for so many years." ■

ARCHAEOLOGY

Restricting use of Indigenous funerary photos splits society

Journal policy aims to boost tribal ties, but critics fear loss of academic freedom

By **Lizzie Wade**

As editor-elect of the journal *Southeastern Archaeology*, Rob Beck helped choose a cover photo for the penultimate issue of 2020. It showed about 20 ceramic vessels, some painted with colorful patterns. They had been excavated in the early 1900s from the Crystal River Archaeological State Park in Florida, home to some of the region's oldest ancient Indigenous earthworks.

But Beck, an archaeologist at the University of Michigan, came to regret his choice. The vessels had been excavated from a funeral mound, and Indigenous members of the Southeastern Archaeological Conference (SEAC) objected, saying the images could severely offend tribal members who viewed them. The journal had already stopped publishing images of human remains for similar reasons. So Beck helped develop a new policy: The journal would publish only line drawings of funerary objects; photographs could only appear in a supplemental online database. And in both cases, researchers would be required to first consult tribes.

But that decision, too, sparked an outcry.

"I felt like I'd been kicked in the gut," says Vin Steponaitis, an archaeologist at the University of North Carolina at Chapel Hill, who says it's crucial for scholars to freely share and study such images. "There'd been no discussion of it with the membership, and I could instantly see that this policy would have a major impact on shutting down research." He was one of 30 SEAC members who signed a petition calling for a vote to revoke the image policy.

Both sides will present arguments on 27 October at SEAC's annual meeting in Chattanooga, Tennessee; the membership will vote in early November.

Carlton Shield Chief Gover, an archaeologist at Indiana University Bloomington

Indigenous scholars objected to publishing photos of funerary objects excavated from an earthen mound complex at Florida's Crystal River Archaeological State Park.

PHOTO: EBYABE/WIKIMEDIA COMMONS

and a citizen of the Pawnee Nation of Oklahoma, says the vote is a key barometer of how much archaeology values collaboration with the descendants of people they study. "Archaeologists across the country watch SEAC," he says.

Beck hopes members will support the new policy. "We can't just bring [Indigenous] folks in and then say, don't change anything." If the vote upholds the restrictions, researchers expect similar policies to spread to other journals, conference talks, and perhaps eventually classrooms. "We have to rethink all of it. And that's not a bad place to be," Beck says.

Avoiding such images is crucial for some Indigenous people. "Funerary objects are sacred and part of the ancestor they were buried with," says RaeLynn Butler, citizen of the Muscogee (Creek) Nation and historic and cultural preservation manager for her tribe. Many tribal members say seeing images of ancestral remains and funerary objects is a violation so profound it can induce physical illness. "From a Muscogee cultural perspective, publishing photos of funerary objects is a form of exploitation," she says.

The point of SEAC's policy is not to uniformly ban all such images, but to encourage collaboration with tribes, says Beau Carroll, lead archaeologist for and member of the Eastern Band of Cherokee Indians and a doctoral student at the University of Tennessee, Knoxville. Along with Beck, he was on the task force that wrote SEAC's image policy. "We're just asking for a conversation that includes us," Carroll says.

Because the U.S. government forcibly removed almost all Indigenous people from the southeast in the 19th century, many of the region's tribes, including the Muscogee (Creek) Nation, now live in Oklahoma. That created the illusion that "it doesn't matter what you do with our remains, because we're gone," Carroll says. But the tribes retain a deep connection to their ancestors and homelands, as well as cultural knowledge that enriches research, he says.

Archaeologists on both sides of the debate agree consulting tribes is an ethical imperative and often improves research. But Steponaitis says scholars ranging from Ph.D. candidates to emeritus professors told him they worried that under the new image policy, their work could no longer be published by their society's flagship journal, and that yearslong projects could be suddenly stalled. What is considered ethical is changing so fast that "the goalposts are moving from week to week," he says.

That's especially concerning to those in the field of iconography, which deciphers the meaning of the motifs decorating funerary objects through detailed visual analysis. Without images of an object, an iconographic study "is unintelligible," says Vernon James Knight, an archaeologist and professor emeritus at the University of Alabama, who signed the petition.

Line drawings are prohibitively expensive and don't capture critical subtleties, and photos published as online supplemental material will likely disappear as technology changes, Steponaitis says. He also argues the new consultation requirement impinges on academic freedom. "If what I publish as a scholar can be suppressed by government officials—and I don't care if it's a state government, the federal government, or a tribal government—that's a huge problem," he says.

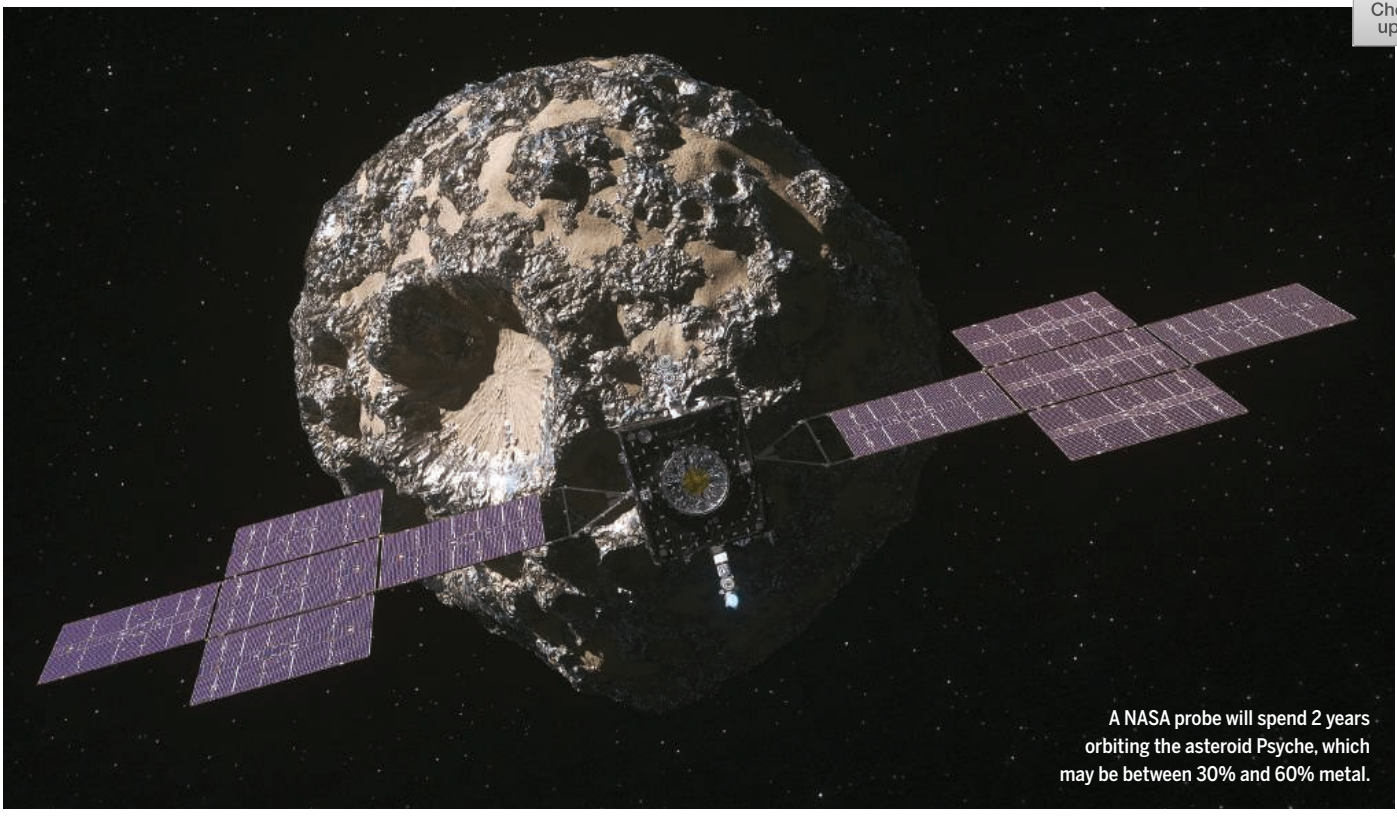
If the restrictions spread, they could hamper education and even preservation, worries Jessica Fleming Crawford, the southeast regional director for the Archaeological Conservancy and a SEAC member who signed the petition. Students who don't see funerary objects in the classroom might not recognize them during salvage work, she says. "I've walked around so many sites surrounded by destruction," she says. "I want us to have the best archaeologists out there working for and with tribes."

Butler and other supporters of the policy compare the consultation requirement to the institutional review board process mandated in other fields of research. They also point out that federally recognized tribes are sovereign nations and have the right to govern their cultural heritage. "It's just like if you were doing archaeology in Greece," where researchers need permits and must comply with heritage laws, says Victor Thompson, an archaeologist at the University of Georgia and a SEAC member who supports the policy.

Carroll warns that if the referendum passes and the image policy is overturned, collaboration "is going to go back 10 or 15 years. It's going to destroy whatever relationship [archaeologists] developed over that time."

The petitioners say their biggest complaint is that SEAC's executive committee approved the policy without putting it to a vote by the membership. Such a significant change "should be a policy the membership is invested in, not one that a small group of members have imposed," Steponaitis says.

The petitioners' concerns about not being consulted are "ironic," Carroll says. "It's pretty much everything Indian people have been saying for 100 years." But Shield Chief Gover, who is currently developing a similar image policy for the Plains Anthropological Society, says archaeologists should have a chance to freely debate their field's changing norms. The approval process in his society, he says, will begin with a membership-wide vote. ■



A NASA probe will spend 2 years orbiting the asteroid Psyche, which may be between 30% and 60% metal.

PLANETARY SCIENCE

NASA launches spacecraft to a mysterious metal-rich asteroid

The asteroid Psyche could be the remnant of a protoplanet, offering insight into the formation of planetary cores

By **Michael Greshko**

When Lindy Elkins-Tanton imagines the metallic asteroid Psyche, she dreams of terrain unlike any seen before. Small craters could look like frozen splashes of water, fringed with silvery spires. Metallic lavas, squeezed out billions of years ago, might shimmer nearby. Gigantic cliff faces, cleaved from the asteroid's metal crust as it cooled and contracted long ago, might be studded with green crystals of olivine.

Whether these vistas exist depends on whether the asteroid, a beguiling 220-kilometer-wide object discovered in 1852, really is the hunk of iron and nickel long assumed by astronomers. Now, Elkins-Tanton and her colleagues are one step closer to finding out, after a $1.2 billion NASA mission to Psyche launched on 13 October. "It's beautifully motivating to be doing a primary kind of exploration to a body that no human has ever seen before," says Elkins-Tanton, principal investigator of the mission, also named Psyche, and a planetary scientist at Arizona State University.

Previous missions to asteroids, which represent the leftovers from planet formation, have explored bodies with mostly rocky exteriors. Psyche will be the first to rendezvous with an "M-type" asteroid: a group of unusually reflective and dense asteroids.

For decades, scientists have wondered whether Psyche is the denuded metal core of a larger protoplanet. According to this scenario, some 4.5 billion years ago, the original body grew massive enough that the heat from its gravitational collapse and radioactive elements partially melted the rocks, which separated into layers. The heaviest metals sank to form a core. Millions of years later, a cataclysmic "hit-and-run" collision with another budding world stripped it of its outermost layers and exposed its iron-nickel heart.

The chance to get a close look at a rocky world's metallic core was a selling point of the mission when NASA selected it in 2017. It promised insights into the composition and history of the deeply buried cores in full-fledged planets. "There's no easy way to probe what a core would be like in Earth or Venus and the rest," says Psyche science team member Bill Bottke, a planetary scientist at the Southwest Research Institute.

Once the spacecraft arrives at Psyche in 2029, it will spend more than 2 years in orbit testing the hypothesis. Subtle shifts in the spacecraft's radio communications with Earth will let researchers map the asteroid's gravity field, revealing whether Psyche is a uniformly dense hunk of metal or glommed together from rubble. A magnetometer will search for relic magnetic fields, perhaps leftover from when the metal was a churning liquid dynamo. Gamma rays and neutrons given off by the asteroid's surface as it is bombarded by the solar wind and cosmic rays should reveal the presence of nickel, expected to make up between 4% and 12% of a metallic core's weight. Imagers will also take pictures of Psyche's terrain that, at their highest resolution, will reveal features as small as basketball courts.

Data from ground-based telescopes have already dealt the metal-core hypoth-

esis some blows. "I'd call that a long shot," says Bloomsburg University of Pennsylvania astronomer Michael Shepard. Updated estimates of Psyche's mass and volume, some from Shepard, have found that it is only about half as dense as iron meteorites. What's more, studies of light reflected from Psyche suggest it isn't purely metal but probably contains some carbonaceous material, a typical asteroid ingredient, and rocky silicate minerals.

The Psyche team now estimates the asteroid is between 30% and 60% metal by volume: not a hunk of pure metal, but perhaps a somewhat porous mix of a protoplanet's core and mantle, mangled by the hit-and-run collision and then given a carbonaceous dusting by billions of years of smaller impacts.

Another possibility is that Psyche has a metallic core hidden by an especially thin, rocky mantle. The shiny patches seen on its surface from afar could be the result of ancient bouts of "ferrovolcanism" that oozed fluid iron sulfide onto the surface. Or Psyche may not have any core or mantle at all. It might just be an unlayered conglomeration of grains assembled in a metal-rich region of the protoplanetary disk of gas and dust surrounding the young Sun.

Rare meteorites called enstatite chondrites have high metal contents and Psyche-like densities, but their parent body is thought to have formed closer to the Sun. The orbits of Psyche and its fellow M-types instead resemble the paths of asteroids that seem to have formed beyond the orbit of Jupiter. "If you're going to make Psyche in all these crazy places, how does it get to the asteroid belt?" Bottke asks.

Getting the Psyche spacecraft to the asteroid belt has also come with challenges. It launched a year late because of slowdowns in software testing that an independent review board partially attributed to understaffing at NASA's Jet Propulsion Laboratory, which manages Psyche. That led to budget overruns that forced NASA to divert funding from VERITAS, a planned Venus mission. "Everyone's doing their best and, you know, sometimes the resources just don't stretch," Elkins-Tanton says.

She hopes for smoother sailing 6 years and 3.6 billion kilometers from now, once the Psyche spacecraft settles into orbit around its namesake. It promises to add to the menagerie of asteroids and comets that humans have explored. The exotic asteroid will also expand our imagination, Elkins-Tanton says: "It's like suddenly discovering there's such a thing as butterflies." ■

Michael Greshko is a science journalist in Washington, D.C.

BIOMEDICINE

Probe of Alzheimer's studies finds 'egregious misconduct'

Co-developer of biotech's drug couldn't supply original data

By **Charles Piller**

Cassava Sciences, a biotech company whose work on the experimental Alzheimer's drug simufilam has been heavily criticized and is the subject of ongoing federal probes, has suffered another blow. A much-anticipated investigation by the City University of New York has accused neuroscientist Hoau-Yan Wang, a CUNY faculty member and longtime Cassava collaborator, of scientific misconduct involving 20 research papers. Many provided key support for simufilam's jump from the lab into clinical studies, and some scientists are now calling for the two ongoing trials to be suspended.

The investigative committee found numerous signs that images were improperly manipulated, for example in a 2012 paper in *The Journal of Neuroscience* that suggested simufilam can blunt the pathological effects of beta amyloid, a protein widely thought to drive Alzheimer's disease. It also concluded that Lindsay Burns, Cassava's senior vice president for neuroscience and a co-author on several of the papers, bears primary or partial responsibility for some of the possible misconduct or scientific errors.

The committee could not prove its suspicions, however, because Wang did not produce the original raw data. Instead, the panel says its finding of wrongdoing was based on "long-standing and egregious misconduct in data management and record keeping by Dr. Wang."

The 50-page report obtained by *Science* says the scientist failed to turn over to the panel "even a single datum or notebook in response to any allegation" and cites "Wang's inability or unwillingness to provide primary research materials to this investigation" as a "deep source of frustration."

Wang's attorney, Jennifer Beidel Scramlin, said neither she nor her client can comment about the report until they speak with CUNY, and that university representatives have not responded to their inquiries. But the report says Wang offered a variety of defenses, including a claim that much of his original data had been "thrown away in response to a request from CCNY [City College of New York] to clean the lab during the COVID-19 pandemic." (CCNY is part of CUNY.) The panel could find no evidence of such a request.

The CUNY investigation into Wang began in the fall of 2021, in response to allegations from other investigators that were forwarded by the Office of Research Integrity (ORI), the federal entity that oversees work funded by the National Institutes of Health. NIH provided millions of dollars for work by Wang and Burns, including about $1.2 million to Wang since December 2020. The investigation was not completed until this year, in May, according to emails obtained by *Science*.

The four-member investigative panel of researchers, led by CUNY neuroscientist Orie Shafer, encountered delays beyond its failed efforts to obtain raw data from Wang. For 6 months, university officials stonewalled the panel's requests for image files confiscated from Wang's computers, according to the report. The investigators received the files only after appealing directly to Vincent Boudreau, president of CCNY.

Science obtained the report, labeled "Final," from a person who requested anonymity because they are not authorized to share it. The university forwarded the report to ORI, according to emails obtained by *Science*. But it has taken no public steps regarding Wang during the roughly 5 months since the report was completed. In response to questions from *Science*, Dee Dee Mozeleski, senior adviser to the CCNY president, said Boudreau could not comment on the matter at this time, but "action on the report is imminent."

CUNY biochemist Kevin Gardner, who helped prepare a preliminary assessment of Wang's work but was not involved in the final review, calls what the panel found "embarrassing beyond words." Wang's record of research is "abhorrent," he says. That this work has supported clinical trials, Gardner adds, "makes it doubly sickening."

In a written statement, Cassava declined to comment about the CUNY report because the school had not shared it with the firm or confirmed to the company the authenticity of what *Science* obtained. It said CUNY did not interview any Cassava employees, "and refused all requests for information and offers of assistance from Cassava." The com-

pany said it "looks forward to continuing the development of simufilam as a potential treatment for patients with Alzheimer's."

Wang and Burns have published together since the mid-2000s. Their studies of the brain's opioid receptors were applied to developing a pain drug that reached clinical trials but never won Food and Drug Administration (FDA) approval. The collaborators shifted their efforts to the possible influence on Alzheimer's of filamin-A, a protein that provides cellular scaffolding for neurons and other cells. Studies conducted in Wang's lab suggested simufilam reinstates the shape and function of filamin-A when it becomes misfolded. In addition, the drug purportedly reduces inflammation and suppresses the toxic effects of beta amyloid and tau, two proteins that can kill nerve cells and have long been implicated in Alzheimer's.

Burns and Wang share inventor credit on numerous patents related to filamin-A and simufilam. They assigned intellectual property rights to Pain Therapeutics—Cassava's name before it was changed in 2019.

Questions about simufilam broke out in public in August 2021, when an attorney filed a petition with FDA on behalf of two neuroscientists—Geoffrey Pitt of Weill Cornell Medicine and David Bredt, a former pharmaceutical executive. The petition called for the agency to pause Cassava's clinical trials. It cited many of the same papers described in the CUNY report, suggesting that apparent image doctoring in them undermined claims that simufilam might slow or even reverse Alzheimer's symptoms.

Pitt and Bredt were short sellers—investors who stand to profit if a company's stock falls. Cassava CEO Remi Barbier challenged their motives and denied their scientific claims, and the company filed a defamation lawsuit against them and other detractors. Meanwhile, Cassava investors filed their own lawsuit claiming that some executives, including Burns and Barbier, who is her spouse, misled investors in a "fraudulent scheme" regarding simufilam's prospects for success.

Cassava conducted relatively small phase 2 trials of the drug. Based on what it said were promising results, it launched the phase 3 trials needed to win FDA approval. One has enrolled more than 800 volunteers with mild-to-moderate Alzheimer's. The company says that by December it will complete enrollment of nearly 1100 more patients for a second phase 3 trial. After the petition, FDA declined to pause those trials, but reserved the right to do so.

Bredt and Pitt's attorney paid Matthew Schrag, a Vanderbilt University neuroscientist, $18,000 to produce some of the analysis used in the petition. Schrag says he has never owned Cassava stock or held a short-selling position and worked independently from Vanderbilt. Schrag sent his findings, along with other apparent irregularities he detected in Wang's papers, to NIH, which passed them along to ORI.

After ORI reached out to CUNY, the school's investigators examined Wang-authored papers published from 2003 through 2021, a conference poster, and a grant proposal to NIH. Many of the 31 allegations the panel reviewed involved apparently improper alterations of Western blots, a technique to distinguish distinct proteins within a tissue sample. Such manipulation can significantly alter the interpretation and validity of experimental findings. The committee said it "found evidence highly suggestive of deliberate scientific misconduct by Dr. Wang for 14 of the 31 allegations."

To check for doctoring of those data, Shafer and his colleagues sought raw-data images to compare against published versions. Wang provided none of those, the report said, adding: "It appears likely that no primary data and no research notebooks pertaining to the 31 allegations exist." The panel also found that Wang "starkly siloed" Western blot preparation in his lab, apparently preparing nearly all such images himself—a highly unusual practice for a lab's principal investigator.

Work by City University of New York neuroscientist Hoau-Yan Wang underlies Cassava's drug simufilam.

Among his defenses, Wang told the investigators that "at least one hard drive" containing key data was destroyed by CCNY officials when they sequestered his materials for review. Wang also accused the committee of bias against him, "failing to follow the CUNY guidelines for this investigation, and of lacking a basic understanding of Western blot analysis." The committee noted in its report, however, that three of its members "routinely conduct experiments involving protein biochemistry and two out of four routinely conduct and publish Western blot experiments."

The CUNY report suggests scientists beyond Wang and Burns might also be faulted. Harvard University neurologist Steven Arnold, senior and corresponding author of a 2012 study with Wang in *The Journal of Clinical Investigation*, "may bear a measure of responsibility" for failing to ensure the integrity of the data in that paper and others co-authored by the CUNY scientist. Arnold leads an Alzheimer's research program at Harvard and, like Wang, has served as a Cassava scientific adviser. Arnold declined to comment, saying he needed time to review the report.

The scientific basis for simufilam, including the critical "misfolding" of filamin-A, came from Wang's lab. But in its statement, Cassava said six basic research studies by organizations that "have no connection to CUNY" also support their drug. One paper, funded by Cassava, concluded simufilam "may promote brain health." Although some of the lab work was done at the Cochin Institute, much of it took place in Wang's lab. Burns and Wang were the chief authors.

Two more papers cited by the company, both co-authored by Burns, suggest simufilam might be effective to treat pituitary tumors or seizures. The other papers examined issues surrounding filamin-A, beta amyloid, and tau, but did not directly test simufilam's ability to reverse or prevent brain pathologies.

The CUNY investigators recommended that journals retract the Wang papers they probed if he can't provide original data. Multiple journals have already done so, after Schrag notified them of his findings. The investigators noted that *PLOS ONE*, which retracted five of Wang's papers, "discovered evidence of research misconduct in his response to the concerns raised." Other journals, including *The Journal of Neuroscience*, posted corrections or expressions of concern on Wang papers, with some editors saying they were waiting for CUNY's investigation before taking further steps.

Wang's fate appears to be in limbo as CUNY prepares its next steps. In the meantime, he continues to publish with Burns, including papers on simufilam in recent months.

Schrag noted that his extensive dossier on Wang's studies went to NIH during the summer of 2021. CUNY's report is "detailed, thorough, and credible," he says. "But a 2-year delay is somewhat astonishing. … The phase 3 simufilam clinical trials should never have started. And they should certainly be shut down on the basis of this report."

This story was supported by the *Science* Fund for Investigative Reporting.

ENERGY POLICY

U.S. hands out $7 billion for hydrogen hubs

Gas could replace fossil fuels and fight climate change—if it is made cleanly

By **Katherine Bourzac**

The Biden administration last week announced $7 billion in funding for seven regional "hubs" to produce hydrogen, which produces water as exhaust when combusted. If made cleanly, hydrogen could help fight global warming by replacing fossil fuels in the fertilizer and steel industries, and in tricky-to-electrify vehicles such as long-haul trucks.

The hubs, which will be funded by the Department of Energy (DOE), are meant to shift hydrogen production away from today's dirty approach: cooking methane under high pressure with steam, a process that has a big carbon footprint. The hubs include both producers and customers, and aim to spur hydrogen adoption by building up storage tanks, pipelines, and other infrastructure.

Hydrogen has had a "chicken and egg" problem, says Keith Wipke, program manager for fuel cell vehicles at DOE's National Renewable Energy Laboratory. "Nobody will start large-scale production until there are customers," he says. And customers are reluctant to switch to hydrogen without a steady, cheap supply. "It's the same story as we've seen with solar and wind. The more you build, the cheaper it becomes," says Anne-Sophie Corbeau, a researcher at the Columbia University Center on Global Energy Policy.

One hub in the Pacific Northwest will focus on so-called green hydrogen, made by using renewable solar or wind power to split water into hydrogen and oxygen. That hub will partner with one in California to develop a hydrogen fueling network, cleaning up power for cargo-handling equipment in ports and preparing for potential hydrogen export. The project will also provide hydrogen to the heavy trucking industry. Today's batteries are too heavy, and charge too slowly, to electrify large trucks. Hydrogen-fueled trucks would not only eliminate carbon emissions, but also other pollutants, improving air quality, particularly for marginalized communities living near trucking centers.

The hubs could help boost demand for electrolyzers, the water-splitting devices that make green hydrogen, which are currently expensive, says Emre Gençer, an energy researcher at the Massachusetts Institute of Technology. The production capacity of electrolyzer manufacturers is low, and buyers often have to wait 2 to 3 years for delivery, he says. "Ideally the hydrogen hubs will create more demand," he says, which will help drive down prices.

The Pacific Northwest hub is the only one that will solely focus on green hydrogen. Others will produce hydrogen from electrolyzers that rely on electricity from nuclear power plants—sometimes called pink hydrogen. Still others will focus on so-called blue hydrogen: made with fossil fuels while capturing the emitted carbon dioxide and reusing or storing it.

An Air Liquide hydrogen plant in Texas. U.S. producers are getting subsidies to make clean hydrogen.

PHOTO: BRETT COOMER/*HOUSTON CHRONICLE* VIA GETTY IMAGES

Environmental advocacy groups are wary of blue hydrogen. Starting with fossil fuels is a problem, says Julie McNamara, deputy policy director in the climate and energy program at the Union of Concerned Scientists. Natural gas fields and pipelines are leaky, she says, and methane is a powerful greenhouse gas. She is also concerned that subsidizing blue hydrogen will entrench the fossil fuel industry and perpetuate the environmental and health impacts that come with it. "This is a chance to set the right path to the future," she says. "We shouldn't be doing hydrogen for hydrogen's sake."

Life-cycle analyses show fossil fuel–based hydrogen production, when paired with carbon capture and storage, can offer climate benefits—but only if it's done with state-of-the-art processes, if the methane supply chain isn't leaky, and if the carbon is efficiently captured and stored, according to a 2022 study by Gençer and colleagues. In the best-case scenario, the study found, blue hydrogen can have a climate impact as low as that of green hydrogen. However, the researchers note that the real-world benefits are likely to be smaller if blue hydrogen production relies on a leaky methane supply chain.

The funding for the hubs comes from the bipartisan infrastructure law. A tax break in a different law, the 2022 Inflation Reduction Act, will also benefit the clean hydrogen industry, especially producers of green hydrogen. The tax credit, available for 10 years, will be awarded to hydrogen producers that meet a low carbon emissions standard. Gençer says this could drive the cost of green hydrogen below $1 per kilogram, which would be cheaper than conventional hydrogen made from methane. Whether blue hydrogen producers can take advantage of the tax break will depend on how the government measures its carbon emissions, and the ultimate cost of their product will depend on the cost of carbon capture and storage.

In a press release, the White House says the awards will spur $40 billion in private investments and create tens of thousands of jobs. Corbeau says the success of the hydrogen hubs will hinge on how well they engage with local communities—in terms of job creation, air quality improvements, and the acceptability of carbon capture and storage. "There have to be positive benefits for people who are going to convert from fossil fuels," she says. ■

Katherine Bourzac is a science journalist based in San Francisco.

AFTER THE FLOOD

As the world's largest dam removal takes shape, restoration ecologists are poised to transform a landscape

By **Warren Cornwall**

Standing on an outcrop of volcanic rock, Joshua Chenoweth looks across the languid waters of California's Iron Gate Reservoir and imagines the transformation in store for the landscape. In early 2024, operators will open the floodgates on the 49-meter-high dam that blocks the Klamath River, allowing the more than 50 million tons of water it impounds to begin to drain. Once it's gone, heavy equipment will dismantle the structure. All that will remain of the 11-kilometer-long reservoir that filled the valley for 60 years will be steep-sided slopes coated in gray mud, split once again by a free-flowing river.

Within months, however, that sediment will be covered with a fine, green carpet of seedlings and colorful splashes of flowers, many planted by Chenoweth's team. Eventually, if all goes as hoped, patches of Gary oak, desert gooseberry, and mock orange will take hold and a lush ribbon of cottonwood, willow, and ash trees will line the banks of the river. Beneath their boughs, salmon that last migrated through this valley more than a century ago will return.

It bears the marks of a mythic creation story—a barren moonscape brought to life. But it's tantalizingly within reach for Chenoweth, a restoration ecologist who has spent his career devising ways to transform

PHOTOS: (LEFT TO RIGHT) PAUL ROBERT WOLF WILSON; JOSHUA CHENOWETH

The Iron Gate Dam (left) is one of four structures slated for demolition along the Klamath River. Scientists hope to restore native ecosystems, such as ponderosa pine woodlands (right), after the reservoirs are drained.

To prepare for a revegetation effort, restoration ecologist Joshua Chenoweth and his team have been growing native plants in greenhouses.

once-submerged ground into thriving forests, meadows, and marshes.

More than a decade ago, he oversaw efforts to restore native ecosystems in Washington's Olympic National Park after two dams were removed along the Elwha River. Chenoweth learned important lessons there. But his current project will be his biggest test yet. Next year, he'll be confronted with nearly 1000 hectares of bare ground stretching along 36 kilometers of the Klamath and its tributaries, after authorities remove Iron Gate and three other dams in California and Oregon.

"It's one of the largest restoration efforts ever," says Chhaya Werner, a Southern Oregon University plant ecologist who plans to study how the vegetation grows back. And with thousands of dams targeted for removal worldwide, more and larger efforts are likely to follow. The Klamath project is an important test of methods that could soon be in wide demand.

Chenoweth, who was hired by the Yurok, a local tribe taking the lead in the revegetation effort, has spent years preparing for the moment. His crews have scoured the surrounding area, diligently collecting seeds that will be used to immediately populate the barren ground. He has enlisted nurseries to grow 260,000 trees and shrubs. And workers have been busy pulling and spraying invasive weeds, hoping to forestall their spread.

But as he gazes out from the rocky promontory, Chenoweth knows success is no sure thing. Invasive medusahead grass, cheatgrass, and yellow starthistle pepper the banks and could quickly spread onto the former lakebed. The unrelenting sun has driven the temperature above 35°C. And the reservoir is the only water in sight. "Getting plants to survive and thrive is always challenging," he says. "In a place like this, it's just harder."

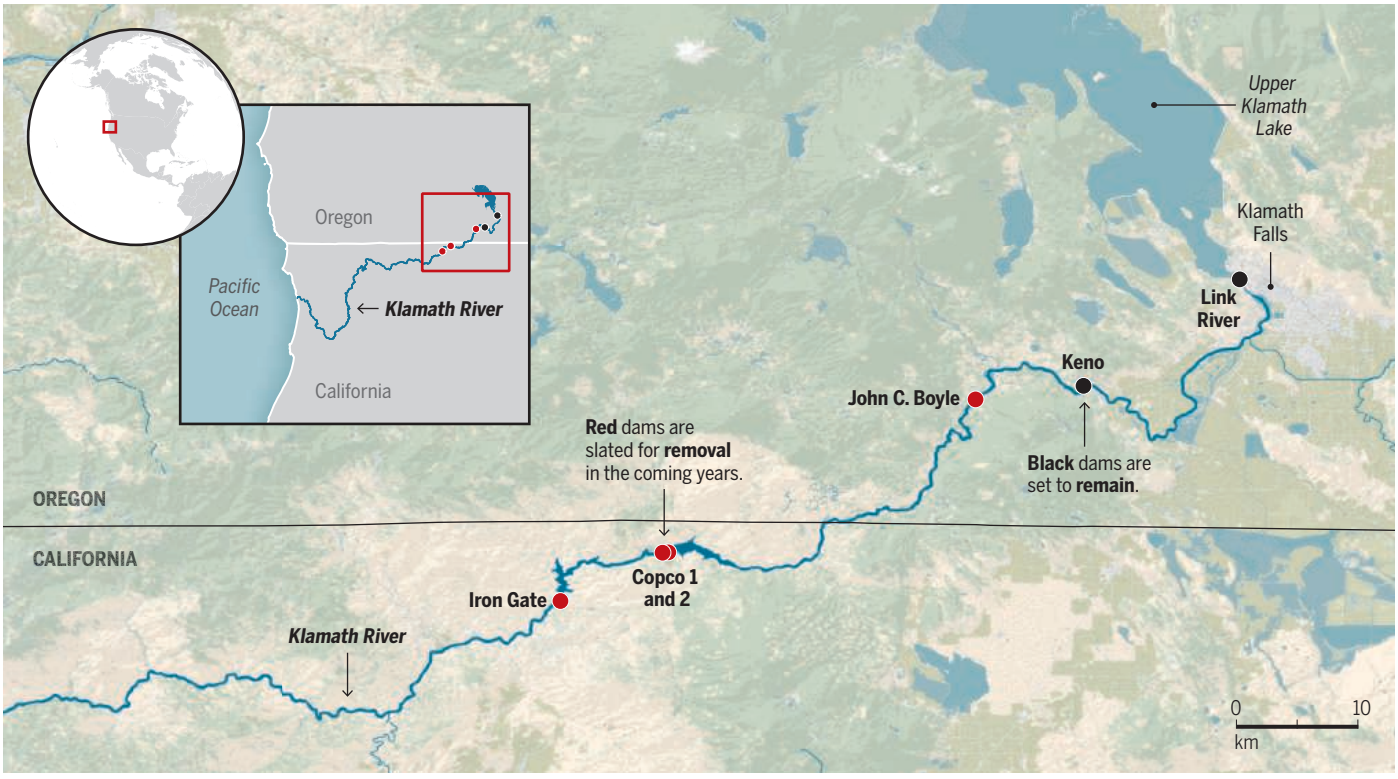
THE KLAMATH WAS ONCE a fabled source of salmon, the third most productive in the western United States after the Columbia and Sacramento rivers. As many as 1 million of the fish journeyed each year up the 420-kilometer-long waterway, which flows from the dry plains of eastern Oregon to the Pacific Ocean. Near the river's mouth, in the lush redwood forests of Northern California, the yearly salmon runs were once the lifeblood of the Yurok, who erected temporary wooden dams to catch them. Farther upstream, as the landscape changes to drier oak and ponderosa forests near the border between California and Oregon, salmon were an important source of sustenance for members of the Karuk, Klamath, and other tribes.

Then, in 1918, an energy company erected a 36-meter-tall concrete dam in a narrow canyon 325 kilometers upriver. With no provision for fish to get past, it effectively walled off some 560 kilometers of habitat for migratory fish on the main river and its tributaries. It was the first of the six dams that now choke the Klamath River. Those blockages—as well as the toxic algae and low-oxygen waters that pooled in the dams' reservoirs—led to severe declines in salmon populations, which now number at less than 5% of predam levels.

In the early 2000s, however, the federal license for many of the dams was approaching its expiration date. Under pressure from tribes, environmentalists, and anglers, the Federal Energy Regulatory Commission signaled that before the license could be renewed, the dams would need to be renovated to help fish get upstream. Facing hundreds of millions of dollars in construction costs, the owner—PacifiCorp—agreed in 2010 to relinquish the dams and allow them to be torn down instead, setting in motion what has become the world's largest dam

Freeing the Klamath

Four large dams along the Klamath River in California and Oregon are scheduled to be torn down by the end of 2024. Although two dams will remain farther upstream, the project is poised to become the world's largest dam removal effort to date and open up more than 600 kilometers of habitat to migratory salmon.



CREDITS: (GRAPHIC) M. HERSHER/*SCIENCE*; (DATA) USGS; AMERICAN RIVERS

removal project, a $450 million to $500 million effort funded by the state of California and PacifiCorp.

The deal was a victory for tribes and environmentalists. "It's about putting things back into balance," says Barry McCovey Jr., a member of the Yurok and director of the tribe's fisheries department.

But it also came with a major conundrum: How do you turn newly exposed land back into something resembling the place it once was? "When we take those dams out and drain those reservoirs, there's going to be a huge scar on the landscape that shouldn't be there," McCovey says.

For insights about how to do that, the tribe turned to Chenoweth and the lessons he learned on the Elwha, another river 700 kilometers to the north.

FOR NEARLY A CENTURY, two dams blocked most of the Elwha, which plunges 72 kilometers from Washington state's glacier-capped Olympic Mountains through coastal forests and into the Pacific. Starting in 2011, the taller of the two dams, a 64-meter-high wall of cement in Glines Canyon, was opened, releasing water that had backed up into a 4-kilometer-long reservoir.

The subsequent removal of that dam, and the smaller one downstream, was at the time the largest project of its kind. So it was difficult to predict what would happen to the landscape as the water levels receded. Scientists had scrutinized how vegetation reclaims places stripped bare by mining or other human activities or by natural forces such as volcanoes and receding glaciers. But the terrain revealed when a dam comes down is different.

After a reservoir is drained, the soil transforms over a single year from a mixture of saturated mud and rough gravel to dry ground. The newly exposed earth differs from what was there historically because dams don't just block water and fish; they also trap fine silt, sand, and rocks. Along the Elwha, the two dams had accumulated enough sediment to fill 8400 Olympic-size swimming pools.

To figure out how to revegetate this unpromising ground, Olympic National Park hired then–36-year-old Chenoweth. Just 6 years earlier, he had quit a career as a TV technical director in New York City. His disillusionment with that work, combined with a passion for plants ignited by a backyard garden in an old Brooklyn farmhouse where he lived, had drawn him across the country to the Pacific Northwest. Now, he was earning a master's degree in restoration ecology at the University of Washington.

Chenoweth and his colleagues weren't sure what kinds of plants could survive in the new soil or what role the surrounding forest could play in seeding it. And the early signs were discouraging. At one point, they scooped silt from the lakebed of the Glines Canyon Reservoir and placed it into planting boxes, where they attempted to grow red alder. Every sapling but one died. "They were really worried nothing was going to grow," says Eastern Washington University plant ecologist Becky Brown, who has tracked how plants have come back along the Elwha.

They stuck to their plan, however. As as soon as water levels started to fall, Chenoweth and colleagues rushed to get the first of more than 300,000 nursery-grown saplings and shrubs into the muck, hoping some would take root. In the end, much of their plan came down to educated guesses and trial and error. "There was nothing I could lean on," Chenoweth says.

Today, looking upstream from the edge of Glines Canyon, it's clear their gamble paid off. Twelve-meter-tall cottonwood trees flash the silvery undersides of their leaves in the stiff breeze, standing on land that

Demolition of the Glines Canyon Dam (left) along the Elwha River was completed in 2014. Today, vegetation is reestablishing on land once flooded by the reservoir (right).

was underwater 12 years ago. Large patches of dark green attest to a welcome surprise: Dense stands of red alder are thriving after all. "Turns out, [the silt experiment was] not a very good analog for nature," Brown notes. "Lots of things grew pretty quickly."

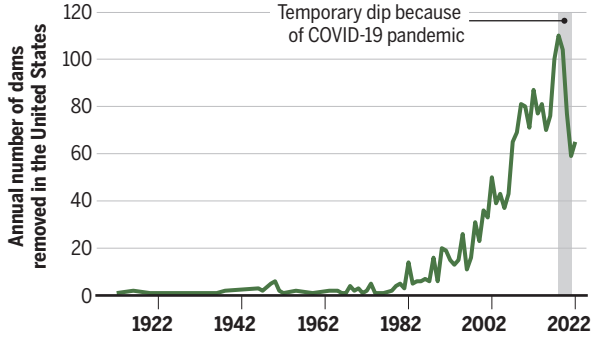
Near the former dam, Brown descends a slope through a cool, verdant forest that reveals key lessons now helping guide work on the Klamath. One is to get plants in the ground as quickly as possible. Moisture left in the soils as the reservoir drained appears to have aided the restoration, creating a brief window for water-loving plants to get a roothold. Of the trees planted in the first year, 82% were still alive 6 years later. By contrast, just one-third of the trees planted in the second year survived.

The thick cottonwood stands are also a reminder to be ready for surprises. There were no efforts to plant this species. Yet, 1 year after the reservoir drained, a 12-hectare forest of cottonwood saplings had unexpectedly begun to emerge. The timing had been lucky: Falling water levels coincided with the spring release of fluffy cottonwood seeds, and because the gravelly soil was still wet when the seeds settled onto it, they survived. Within a few years, there were 9600 young cottonwoods in a single hectare.

Also notable is what is not growing. The invasive species that dominate many disturbed landscapes in western Washington—scotch broom, Himalayan blackberry, Canada thistle—haven't taken over. "The assumption … is that invasive species will become the dominant plant in these dam removal projects," Chenoweth says. But on the Elwha, that hasn't turned out to be the case. In part, Chenoweth credits work to battle weeds by mowing and spraying herbicide at key locations, as well as a relative scarcity of invasives inside the national park. But the spreading of native seeds, he says, also helped. In research plots that weren't seeded, 27% of the plants weren't native. In places that were seeded, that percentage fell to 19%.

Rising demolitions

The United States leads the world in dam removal projects. Demolitions surged in the past 2 decades, peaking at more than 100 per year before the COVID-19 pandemic slowed work.



THE KLAMATH MAY BE a harsher test. The dearth of water and omnipresent weeds have Chenoweth on edge and taking a more assertive stance.

Alert to the need to get plants in the ground quickly, he has directed a yearslong effort to build a seed bank of more than 90 different species, totaling nearly 30,000 kilograms of seeds. Many were produced at nurseries capable of turning a small stash of wild seeds into a bounty over several years. But some plants don't thrive in nurseries. For those, Chenoweth is relying on crews to gather the seeds by hand.

On a sunny June morning, one of those teams stands on a hillside

that slopes into a bay on Iron Gate Reservoir, the westernmost of three water bodies on the Klamath that are slated to disappear. Alauna Grant, a member of the Karuk Tribe and the crew leader, slowly circles a scraggly 2-meter-tall antelope bitterbrush, deftly plucking tiny seeds no bigger than grains of wheat. "This one is so important, because it thrives in this habitat," she says.

Grant has spent four summers collecting seeds from native plants in the region. It's difficult work. After half an hour, her freshly picked seeds barely cover the bottom of a small paper bag and the temperature is already above 35°C, with little shade. But she says she and her crew, members of the Hupa and Yurok tribes, relish the chance to work outside and contribute to rebuilding the river. "We're doing something that matters," she says.

Nearby, another group is working to remove unwanted plants and create weed-free strips at critical places along the edge of the reservoir, such as where creeks will eventually feed into the free-flowing river. Richard Green, a California State Polytechnic University, Humboldt tribal forestry student and member of the Yurok, crouches and wraps his gloved hand around a half-meter-tall yellow starthistle. Leaning back, he pulls it from the ground, adding it to a growing bundle of uprooted plants. With its spiky macelike flowers and seeds that cling to any passing creature, the plant could quickly overrun the freshly exposed ground. "They're no joke," Green says, with grudging respect.

His hand-to-stem combat with the weed is part of a larger—and deeply personal—mission for Green. His grandmother, Bonnie Green, was a longtime political leader of the tribe before she died in 2014. He recalls the years she spent traveling to Sacramento, California, and Washington, D.C., lobbying for the dams to come down.

Chenoweth is now training Green to help manage the vegetation program in the future. It's key to hopes for a healthy river. Verdant banks will help keep the water cool and control erosion, which could influence whether populations of temperature-sensitive salmon are able to rebound. "I don't feel [the pressure] in a stressed out way," Green says as he pauses from the weeding. "It feels like a good urgency to keep it going."

ONCE PORTRAYED as environmental fantasies, dam removals such as the ones planned along the Klamath are increasingly gaining traction as structures erected during the 20th century age, awareness of their ecological toll grows, and the list of demolished dams lengthens. "Twenty years ago, everybody thought the Klamath [dam removal] was not even a moonshot. It was just a nonstarter," says Ann Willis, head of the California office of the environmental group American Rivers.

Willis's group, a leading proponent of dam removals, has a goal of seeing 30,000 U.S. dams torn down by 2050, a dramatic increase over the roughly 2000 structures taken down over the past century, according to an American Rivers database. Many barriers are so small that their removal would go unnoticed by the public. Others, however, are river-blocking mammoths. In California, top candidates include two dams on the Eel River, south of the Klamath, and a string of six dams on Battle Creek, a tributary of the Sacramento River. Farther north, four huge federally owned dams on the Snake River along the border between Idaho and Oregon are the biggest targets in the country and the subject of a decadeslong political battle.

Alauna Grant, a member of the Karuk Tribe, has spent four summers collecting seeds in the Klamath River region to plant after the dam removals.

Dam removal is picking up steam in Europe as well, where more than 300 dams, culverts, and other river-blocking stuctures were removed in 2022, according to Dam Removal Europe, a coalition of environmental groups. In July, the European Parliament approved a draft of a nature restoration law with a target of making at least 20,000 kilometers of rivers free flowing by 2030. If the law is enacted, "all European countries [will] have to start thinking about this," says Herman Wanningen, director of the Dutch-based World Fish Migration Foundation.

Each removal would launch another restoration experiment, sometimes on a grand scale. The removal of the Snake River dams, for instance, would expose an area equal to Manhattan Island along 225 kilometers of river. Willis sees the Klamath project as vital to predicting how other river ecosystems may bounce back, especially ones that are less pristine than the national park that surrounds the Elwha.

But it will be years before the fate of the new Klamath landscape becomes clear. Even on the Elwha, Brown and her students continue to track how plants are reclaiming the terrain. One stretch of gravel still bears few signs of life, save some tiny wild strawberries that stretch pink tendrils across the ground. Walk 30 meters, and a patch of 10 Douglas fir trees no taller than teacups sprout from a mosaic of gray stones. "The Doug fir … is getting here on its own, which is really cool," she says. "We'll see how it looks in 5 years."

As work proceeds on the Klamath, Chenoweth plans to track the progress of revegetation efforts at 181 monitoring spots for 5 years—some planted and others left to the whims of the environment. Knowing what thrives and what doesn't will enable him to adapt and potentially change course as the revegetation effort proceeds.

In parallel, Werner plans to monitor how key plants, including Gary oak, buckbrush, California poppy, and yarrow, respond to different planting densities on nearly 100 plots. She'll also be tracking whether invasive plants are able to gain a roothold and how environmental factors such as water and sun exposure influence the vegetation community. The project, she says, could answer "big picture ecology questions" such as what determines winners and losers in the fight for open ground.

It will be illuminating to see how that plays out, Chenoweth notes, because for all of the sweat, planning, and research, there is only so much humans can do to shape the landscape. Back at the rocky overlook along the Iron Gate Reservoir, he contemplates the forces beyond his control, wondering whether the new seedlings planted along this stretch of the Klamath will get a boost from a wet spring or face a punishing drought. "Mother Nature is going to do the hard work," he says. "But if Mother Nature fails, it's going to be really hard for us." ■

PHOTO: W. CORNWALL/*SCIENCE*

ECOLOGY

Are crabs in hot water?

Marine heat waves challenge sustainable fisheries management

By **Gordon H. Kruse**

An estimated 10 billion snow crabs (*Chionoecetes opilio*) disappeared from the eastern Bering Sea off western Alaska, USA, during 2018 to 2021. This loss encompassed more than 90% of the Bering Sea snow crab population and nearly four times the combined total number of snow crabs harvested by the fishery from 1977 to 2022 (*1*). Before this event, the population was healthy, with a huge incoming group of immature crab. On page 306 of this issue, Szuwalski *et al.* (*2*) use a population model, statistical analyses, and various lines of evidence to investigate reasons for this loss of crab. They concluded that starvation associated with a marine heat wave and a high density of mature crab was the likely culprit. This finding highlights the importance of using ecosystem information in fisheries management. However, performing data collection, analysis, and synthesis fast enough to provide definitive management advice in real time is challenging.

Fisheries off the coast of Alaska are generally considered to be well managed because of broad use of scientific assessments of abundance, low harvest rates (*3*), and other conservation measures. However, crab fisheries off of Alaska have a volatile history. Typically, fluctuations in Alaska crab populations are attributed to wide variations in the numbers of young crabs produced each year. This variation occurs regardless of the number of adults, suggesting that it is controlled by environmental factors (*4*). However, there have been a few instances in which mass die-offs of crabs of both sexes and all sizes, not a lack of immature crabs, appear to have led to sharp declines in the populations of Bering Sea crab: Bristol Bay red king crab (*Paralithodes camtschaticus*) in the early 1980s (*5*) and St. Matthew Island blue king crab (*Paralithodes platypus*) in the late 1990s (*6*). A lack of data prevented detailed analysis of potential causes of these earlier die-offs.

The conclusion by Szuwalski *et al.* that starvation is the likely cause of the mass mortality observed in 2018 to 2021 is consistent with several lines of evidence. Snow crab is an Arctic species whose distribution has been shifting north for decades as the Bering Sea becomes warmer (*7*). The east-

Snow crabs (*Chionoecetes opilio*) are an important catch for Bering Sea fishing communities.

ern Bering Sea experienced anomalously warm conditions in 2018, broke temperature records in 2019, and had an unprecedented near-complete lack of sea ice in both years. Caloric requirements of laboratory-reared snow crab nearly double as water temperature increases from 0° to 3°C (*8*). Moreover, during this marine heat wave, the footprint of snow crab was among the smallest observed in the history of Bering Sea surveys (which began in 1975), suggesting overcrowding. Statistical analyses by Szuwalksi *et al.* found that mortality of both immature and mature crabs was related to temperature and density of mature crabs. So, it seems plausible that higher caloric requirements and more competition led to starvation.

Further work is required to establish why the mortality of immature crabs was greatest at intermediate densities of mature crabs, whereas mortality of mature crabs was greatest when mature crab density was highest. Moreover, Szuwalski *et al.* used reduced weight relative to body size as evidence of starvation, but snow crabs starved in the laboratory do not lose weight (*9*), and their weight also varies with other factors, such as molt cycle (the process of shedding their external shell so that they can grow in body size). Starvation may be better indicated by measures of lipid storage or fatty acids (*10*).

The 2018 to 2019 marine heat wave also had broad implications throughout the Bering Sea ecosystem, including for phytoplankton, zooplankton, fish, seabirds, and marine mammals (*11*). Likewise, a 2014 to 2016 marine heat wave led to abrupt changes throughout the Gulf of Alaska ecosystem (*12*). Notably, the number of commercially important Pacific cod (*Gadus macrocephalus*) declined by 71% from 2015 to 2017, and their die-off was explained by an increase in metabolic demand and reduced availability of prey associated with the heat wave (*13*). Such marine heat waves are likely to become more frequent and severe with global warming (*14*).

These vast ecological changes also have social and economic costs. Dockside payments to fishermen for snow crab declined from US$133 million in the 2020–2021 fishing season to US$24 million in 2021–2022, to nil in 2022–2023 (when the fishery closed) (*1*). The fishery collapse is having severe economic impacts on stakeholders in Alaska and the US Pacific Northwest, including remote fishing-dependent communities. The Bering Sea snow crab fishery was declared a disaster in 2022 by the US Department of Congress, and hundreds of millions of dollars were awarded to assist members of affected fishing communities, including fishermen, crew members, and processors.

The study by Szuwalski *et al.* demonstrates that marine heat waves, superimposed on long-term trends of ocean warming, can lead to major and sudden ecosystem changes with severe economic consequences. It highlights a need for ecosystem-based fisheries management, a holistic approach that recognizes all the interactions within an ecosystem rather than considering a single species or issue in isolation. The challenge is to quantify cause-and-effect relationships between ecosystem changes (for example, in temperature or predator abundance) and resultant changes in processes that affect the distribution and productivity of a harvested species (for example, reproductive success, mortality, and movement). In the Alaska region, the North Pacific Fishery Management Council has taken the first step toward this aim by developing a standardized framework, called the ecosystem and socioeconomic profile, that compiles and evaluates relevant indicators as part of the stock assessment process (*15*). The resilience of fisheries to climate change will depend in part on the ability of science to anticipate and communicate ecosystem changes to fishing communities in advance so that they can reallocate resources and effort to minimize economic impacts.

Providing actionable scientific advice is challenging in the early stages of unusual ecosystem events (such as marine heat waves) that are being observed for the first time. In such cases, fishery management often defaults to a reactionary mode, acting after major effects such as a population collapse have already been realized. So taking an ecosystem-based fisheries management approach, if a similar marine heat wave scenario with snow crab occurs sometime in the future, what should the response look like? Some stakeholders argue that harvest rates should be increased as the stock declines so that the crabs can be caught before they die. An alternative approach would be to decrease harvest rates in an attempt to maintain some critical abundance of adult crabs that can successfully reproduce. Lessons can be learned from a major late-1970s climate-driven ecosystem shift that saw increases in some marine species and declines in others, including red king crab throughout the Gulf of Alaska, Aleutian Islands, and eastern Bering Sea. Some of the crab stocks were subjected to high harvest rates before the fisheries were closed, and 40 years later, few red king crab populations have recovered. ■

College of Fisheries and Ocean Sciences, University of Alaska Fairbanks, Juneau, AK, USA. Email: ghkruse@alaska.edu

REFERENCES AND NOTES

1. E. Nichols, J. Shaishnikoff, "Fishery management report 22-28," Alaska Department of Fish and Game (2022).
2. C. S. Szuwalski *et al.*, *Science* **382**, 306 (2023).
3. R. Hilborn *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 2218 (2021).
4. J. Zheng, G. H. Kruse, *Prog. Oceanogr.* **68**, 184 (2006).
5. K. J. Palof, M. S. M. Siddeek, "Bristol Bay red king crab stock assessment in Fall 2022," North Pacific Fishery Management Council (2022).
6. K. Palof, J. Zheng, J. Ianelli, "Saint Matthew Island blue king crab stock assessment," North Pacific Fishery Management Council (2020).
7. F. J. Mueter, M. A. Litzow, *Eco. App.* **18**, 309 (2008).
8. T. P. Foyle, R. K. O'Dor, *J. Exp. Biol.* **145**, 371 (1989).
9. S. L. Siikavuopio *et al.*, *Aquacult. Res.* **50**, 550 (2019).
10. L. A. Copeman *et al.*, *Polar Biol.* **44**, 1883 (2021).
11. E. C. Siddon, S. G. Zador, G. L. Hunt Jr., *Deep-Sea Res. II* **181-182**, 104914 (2020).
12. R. M. Suryan *et al.*, *Sci. Rep.* **11**, 6235 (2021).
13. S. J. Barbeaux, K. Holsman, S. Zador, *Front. Mar. Sci.* **7**, 703 (2020).
14. T. L. Frölicher, E. M. Fischer, N. Gruber, *Nature* **560**, 360 (2018).
15. S. K. Shotwell *et al.*, *Deep-Sea Res. II* **198**, 105070 (2022).

ACKNOWLEDGMENTS

G.H.K. serves as a Science Panel member for the North Pacific Research Board and Exxon Valdez Oil Spill Trustee Council and as science adviser to the Bering Sea Fisheries Research Foundation.

10.1126/science.adk7565

NEUROSCIENCE

Cracking the neuronal code of episodic memory

Hierarchical organization of memory is observed in the brains of rats

By **Jasmin S. Steudler** and **H. Freyja Ólafsdóttir**

Episodic memory is the ability to associate memories of events with their spatial and temporal contexts (*1*). Intrinsic to episodic memory is the ability to draw on past experiences to make predictions about the future (*2*). For example, while searching for food rewards in a familiar maze, a rodent can predict upcoming reward locations based on the memory of past trials. Research over the past decades has shown that the hippocampus, and the location-locked activity of hippocampal principal neurons (place cells), is critical for the associative and predictive aspects of episodic memory (*3*, *4*). Yet, whether these two distinct functions are supported by common or dissociable neuronal codes has remained unclear. On page 283 of this issue, Liu *et al.* (*5*) report that hippocampal associative and predictive codes in rats are supported by interrelated but distinct circuit mechanisms, suggesting that episodic memory may be hierarchically organized in the brain.

Hippocampal place cells, typically observed in mice and rats, are neurons that display location-specific activity that encodes relationships between a spatial location and environmental landmarks and/or boundaries (*3*, *6*). Place cells that are active in the same spatial location are said to belong to the same cell assembly (*7*). As such, the location specific activity expressed by cell assemblies may represent the neuronal substrate of associative memory.

Place cells also express a sequence code: Nested within individual cycles of theta-band oscillations—oscillations in neuronal activity observed in the rodent hippocampus during locomotion—are multiple cell assemblies that fire in sequence and encode relationships between sequentially visited places (theta sequences). Notably, during theta sequences, the spatial representation that cell assemblies depict proceeds ahead of the animal. In other words, theta sequences provide a prediction for upcoming locations (*8*). These putative associative (assembly) and predictive (sequence) hippocampal codes are then replayed during sharp-wave ripple (SWR) oscillations, which are observed during periods of sleep or rest, thereby committing them to long-term storage (*9*, *10*). Although these hippocampal coding schemes have been studied extensively, their contribution to specific aspects of episodic memory has not been demonstrated.

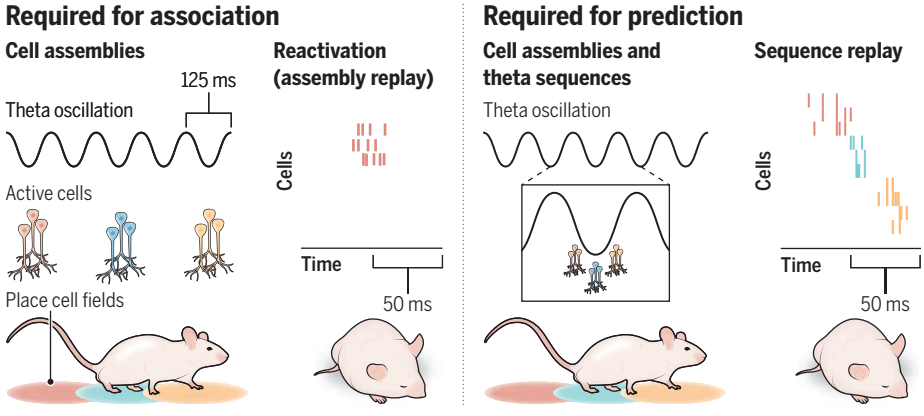
To address this gap in knowledge, Liu *et al.* used an optogenetic approach to control the activity of γ-aminobutyric acid (GABA)–expressing neurons in the medial entorhinal cortex (MEC) of rats. These neurons provide the main input to the hippocampus and control the timing of the activity of principal hippocampal cells in relation to theta-band oscillations. By entraining the GABA-expressing cells to an artificial frequency, the authors were able to abolish theta sequences while sparing the expression of hippocampal cell assemblies. Disrupting MEC input to the hippocampus also led to a selective disruption of theta sequence replay, whereas the replay of cell assemblies was unaffected. Therefore, the authors were able to disrupt the putative predictive code of the hippocampus while keeping the associative code intact.

Liu *et al.* assessed how this disruption affected learning in two paradigms that rely on different episodic memory processes. In one paradigm, rats learned to associate one half of a circular platform with reward [conditioned place preference (CPP)], and in another, the animals learned to sequentially navigate to three reward locations on the platform. The former task is a test of associative memory because it requires animals to learn an association between a particular area and reward, whereas the latter task may additionally require predictive capabilities because optimal navigation in this task requires the rat to learn how different places relate to each other—that is, to predict the location of the next reward. The authors found that disrupting the putative predictive code did not impair learning in the CPP task. However, performance in the spatial navigation task was severely affected, indicating that the associative code is sufficient for CPP learning but is insufficient for spatial navigation learning (see the figure).

These observations indicate that consolidation of the place cell predictive code by replay sequences may be required for at least some aspects of episodic memory function. Multiple studies have shown that disrupting SWRs and cell assembly replay interferes with learning (*10*, *11*). However, direct evidence that sequence replay supports aspects of episodic memory has been lacking. Of equal importance is the finding that the predictive code may be dispensable for other episodic memory processes. Liu *et al.* found

Encoding of association and prediction

The associative aspect of episodic memory in rats requires the firing of location-specific assemblies (indicated by color) of place cells in the hippocampus but does not need this firing to be synchronized with theta oscillations. The reactivation of these assemblies during sleep or rest, which consolidates the memory, is also required. The predictive aspect of episodic memory requires the place cell assemblies to fire in sequence within theta oscillations and for this sequential firing to be replayed during sleep or rest.



Donders Centre for Neuroscience, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands. Email: freyja.olafsdottir@donders.ru.nl

that rats could form a long-term associative memory in the absence of theta and replay sequences; this suggests that the expression of cell assemblies and their later reactivation during rest periods may be sufficient for core component processes of episodic memory.

These findings have important implications for understanding neurodegenerative and neurodevelopmental disorders that affect episodic memory, including Alzheimer's disease and autism spectrum disorder. Finding that associative memory and predictive capabilities may be supported by dissociable neuronal codes could mean that these specific facets of episodic memory are differentially affected in these disorders and/or are affected at different stages of a progressive disorder, such as Alzheimer's disease. Hopefully, future research will investigate associative and predictive capabilities in neurological conditions.

Additionally, the findings by Liu *et al.* raise questions about whether the putative hippocampal predictive code holds for other tasks that draw on predictive capabilities. One interpretation of the result is that the predictive code (expressed through sequential hippocampal activity) was required for the spatial navigation task because of its sequential nature. Perhaps future studies can assess whether the hippocampal predictive code is also involved in more generic prediction learning. For example, does it help guide navigation when barriers are introduced into a familiar environment?

Investigating the role of the predictive code in inferential reasoning (inferring a relationship between two different events) is also an important research avenue. The ability to use spatiotemporal information to make inferences is central to predictive capabilities, as noted by Liu *et al.* and others (*2*, *4*). Does the described hippocampal predictive code guide learning when, for example, an animal needs to find reward in a previously unseen environment that is similar to a familiar environment? Identifying the neuronal coding scheme supporting such inferential reasoning represents an exciting research frontier. ■

REFERENCES AND NOTES

1. E. Tulving, in *Organization of Memory*, E. Tulving, W. Donaldson, Eds. (Academic Press, 1972), pp. 381–403.
2. E. Tulving, *Annu. Rev. Psychol.* **53**, 1 (2002).
3. J. O'Keefe, L. Nadel, *The Hippocampus as a Cognitive Map* (Clarendon Press, 1978).
4. H. Eichenbaum, N. J. Cohen, *Neuron* **83**, 764 (2014).
5. C. Liu *et al.*, *Science* **382**, eadi8237 (2023).
6. N. Burgess, J. O'Keefe, *Hippocampus* **6**, 749 (1996).
7. K. D. Harris *et al.*, *Nature* **424**, 552 (2003).
8. A. Johnson, A. D. Redish, *J. Neurosci.* **27**, 12176 (2007).
9. D. J. Foster, M. A. Wilson, *Nature* **440**, 680 (2006).
10. G. Girardeau *et al.*, *Nat. Neurosci.* **12**, 1222 (2009).
11. I. Gridchyn *et al.*, *Neuron* **106**, 291 (2020).

10.1126/science.adk4642

COMPUTER SCIENCE

AI computing reaches for the edge

A chip design integrates computation and memory to efficiently process data at low energy cost

By **Subramanian S. Iyer** *and* **Vwani Roychowdhury**

Artificial intelligence (AI)—the ability of computers to perform human cognitive functions in real-world scenarios—requires substantial computation power, energy, and vast datasets. Once trained, AI models are deployed to make predictions (inferences) for new situations. In the traditional paradigm, inference is supported by centralized, high-performance computational platforms and high-bandwidth network connections. However, in real-time mission-critical applications such as facial recognition, object detection tracking, and behavior monitoring, an "edge" computing system is desirable, for fast and accurate inference and, hence, fast response times. Edge computing requires moving the large AI model from a centralized location to a position closer to the source of data (hence, working at the edge). On page 329 of this issue, Modha *et al.* (*1*) describe a computing platform called "NorthPole" that facilitates high inference speed and prediction accuracy but with a moderate energy requirement. This is a promising step toward chip designs that support low-power edge AI inference.

"...a promising step toward chip designs that support low-power edge AI inference."

Learning integrates immense sets of experiential data that are located in large, centralized data centers. It is an energy-, computing-, memory-, and time-intensive set of operations that results in a trained network with learned parameters (such as a weight value of input data). The trained model often comprises billions of parameters. Thus, the inference task is computationally demanding and primarily consists of matrix-vector multiplication operations, referred to as multiply and accumulate computations (MACs). Devices "at the edge" that monitor and track in real time, for example, must communicate with powerful centralized AI models and then bring information back to the edge. This route is fraught with bandwidth limitations, latency problems, and potential disruption to networks. The objective of edge computing is to bring the large AI model to the edge, to accelerate inference. However, this requires a system at the edge that supports computation while simultaneously accessing data—much like the human brain.

Modha *et al.* designed a brain-inspired computer chip architecture that focuses on efficiently performing inference tasks with technological, algorithmic, and software innovations. The platform uses a two-dimensional array of cores (computing units or, simply, compute units) and associated memory blocks that store weights of AI models. The MACs required for inferencing are assigned to different compute cores, each facilitating the multiplication of the weights of an AI model against input data. This requires that necessary weights be routed to the appropriate cores on demand, a high energy-consuming task. One innovation of NorthPole is the ability to share compute and memory (weights) across the chip with no or minimal hierarchy. This means that a core compute unit can access memory that is both spatially close as well as far across the chip with almost equal ease. Another notable achievement by Modha *et al.* is that the chip, architecture, and software are co-optimized for edge inference and adapted to standard semiconductor technology for manufacturing integrated circuits. Because NorthPole is a digital system, it also has the advantage of being resistant to device noise and systemic biases and drifts that afflict analog systems. The primary source of inference error, however, is the use of limited-precision (number of bits) arithmetic operations for MACs. Nevertheless, NorthPole's 256 cores and 224 MB of on-chip memory can operate

Electrical and Computer Engineering Department, Samueli School of Engineering, University of California, Los Angeles, CA, USA. Email: s.s.iyer@ucla.edu; vwani@ucla.edu

at a frequency range of 25 to 425 MHz, with an energy consumption of >1000 frames per joule, making it suitable for applications that include a self-driving car.

A limitation of Modha *et al.*'s approach is its relatively high power requirements when scaled to large AI models. This is partly due to the data architecture (how data are collected, stored, and flowed through the system) that permits low-latency inference by making data transfers possible from memory located both near and far across the chip. Increased power and size demands also arise from elaborate peripheral chips (such as field-programmable gate arrays coprocessors) and memory that are needed to power up the compute-memory integrated chip from a cold start. From this perspective, an edge device using NorthPole must have access to a reasonably high-power source and incur any associated increase in its size. Several applications of AI inference chips, however, include functioning within sufficiently powerful machines such as automobiles, aircraft, and military vehicles, which would benefit from the chip design of Modha *et al.*

"...analog approaches may ultimately bring real-time AI to hand-held edge devices..."

In the quest for lower-power AI edge inference accelerators, a somewhat different architecture called compute-in-memory (CIM) has been explored (*2*). CIM brings memory (weights) and compute (multipliers) units physically close together, at the level of the multiplier units used in MACs (recall that MACs are the primary operations required for inference). CIM can be either digital or analog. In digital CIMs, the weights are stored as bits using static random-access memory (which retains data as long as power is being supplied) and the multiplication is finite precision binary arithmetic, similar to the operation used by Modha *et al.* An analog CIM is more elegant in that it directly uses Ohm's law to perform multiplication and uses a capacitor as an adder. It has potentially much lower power needs compared to digital MACs. The weights are stored as the conductance of a conductive element, and current flowing through the element is proportional to the product of input (applied voltage) and the conductance (weights). The currents through several such elements are added to provide the primary inference functionality. In principle, this MAC operation is very fast, and the power for inference is attractively low.

The challenge, especially in the analog CIM approach, lies in the choice of the conductive element and the resolution of its conductance, which determines the resolution of the weights. In one example, a combination of several phase-change memory elements is used to store the weights (*3*). However, such analog CIM systems require the integration of new materials, which necessitates the addition of specialized steps to the industry-standard manufacturing process for complementary metal oxide semiconductors. An approach that addresses this obstacle requires instead a standard semiconductor device that traps charge (called the charge trap transistor) and uses it to tune its resistance and achieve an estimated 8 equivalent bits of resolution in the resulting weights (*4*).

A disadvantage of analog techniques is that programming the weights is cumbersome, time consuming, and energy intensive. Thus, more suitable for edge applications would be AI models that are updated only occasionally. In both analog CIM and digital approaches, precision of the weights and computations play a key role in accelerating inference. Thus, more work is necessary on designing training algorithms to yield AI models that are robust to both reduction in precision and systemic biases in analog devices. Recent progress reported for analog (*2*) and digital (*5*) platforms is encouraging.

Innovative chip architecture and digital approaches such as those reported by Modha *et al.* will play an important role in the near-term development of edge-based inference with moderate-size platforms. However, analog approaches may ultimately bring real-time AI to hand-held edge devices, especially as the enabling technologies become more mainstream and easily manufacturable. ■

REFERENCES AND NOTES

1. D. S. Modha, F. Akopyan, A. Andreopoulos, R. Appuswamy, J. V. Arthur, A. S. Cassidy, P. Datta, M. V. DeBole, S. K. Esser, C. Ortega Otero, J. Sawada, B. Taba, A. Amir, D. Bablani, P. J. Carlson, M. D. Flickner, R. Gandhasri, G. J. Garreau, M. Ito, J. L. Klamo, J. A. Kusnitz, N. J. McClatchey, J. L. McKinstry, Y. Nakamura, T. K. Nayak, W. P. Risk, K. Schleupen, B. Shaw, J. Sivagnaname, D. F. Smith, I. Terrizzano, T. Ueda, *Science* **382**, 329 (2023).
2. Z. Wan, T. Wang, Y. Zhou, S. S. Iyer, V. P. Roychowdhury, *ACM J. Emerg. Technol. Comput. Syst.* **18**, 1 (2022).
3. M. Le Gallo *et al.*, *Nat. Electron.* **6**, 680 (2023).
4. X. Gu, Z. Wan, S. S. Iyer, *IEEE Trans. Electron. Devices* **66**, 10, 4183 (2019)
5. T. Dettmers, M. Lewis, S. Shleifer, L. Zettlemoyer, "8-bit optimizers via block-wise quantization," 9th International Conference on Learning Representations, ICLR 2022, 25 to 29 April 2022.

10.1126/science.adk6874

PHOTONICS

Tracking light-induced charge transport

Precise charge dynamics could help to improve the operation of solar cells and sensors

By **Rachel E. Bangle** *and* **Maiken H. Mikkelsen**

Key to developing photocatalytic and solar cells is understanding charge dynamics. Light absorption in metals can initiate charge transport by exciting electrons to form high-energy "hot" charges, but this energy is rapidly lost as heat (*1*). To capture the extra energy that hot charges (or carriers) possess before they decay and use it to generate an electrical current, the carriers must be transported across an interface called a transport junction. The movement of hot charges determines the efficiency of these junctions. However, hot-carrier dynamics are notoriously difficult to measure owing to their ultrashort lifetimes (*2*–*4*). On page 299 of this issue, Taghinejad *et al.* (*5*) show that ejection of hot carriers at a transport junction produce terahertz (THz, 10^{12} Hz) radiation that provides information exclusively on hot-carrier dynamics. This could provide a reliable way to characterize hot carriers and thereby aid the design of sensors and solar cells.

Hot carriers are formed when collectively oscillating electrons, called plasmons, transfer energy within the metal to form excited electron and hole (positive charge) states on timescales of <100 fs (1 fs = 10^{-15} s). This initial state is very energetic and thus outside of the usual thermal equilibrium of electrons in metal. However, nonequilibrium carriers rapidly (<1 ps, 1 ps = 10^{-12} s) spread that energy among a larger electron population through scattering collisions. This raises the average electron temperature, and the electrons quickly (<100 ps) dissipate the excess energy to the metal lattice as heat. Hot-carrier extraction can occur during

Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA.
Email: m.mikkelsen@duke.edu

this process if the electrons have enough energy to overcome the energetic barrier at a metal-dielectric interface (the transport junction). It is unknown, however, whether electrons in thermal equilibrium and excited through scattering can participate in charge transport along with nonequilibrium carriers (*6*).

Experiments directed at measuring hot-carrier dynamics are obscured by the multiple hot-carrier decay pathways that occur on similar timescales in the same locations (*7*). Traditional charge transport measurements have reported numbers of photoejected electrons (*8*) and have quantified the energetic distribution of hot carriers (*9*). These measurements, however, have not provided insights into hot-carrier dynamics. Ultrafast optical measurements have shown electron ejection to occur within 50 fs but have been unable to distinguish at which stage in the energetic decay hot carriers undergo ejection (*10*). Similar measurements have elucidated subpicosecond plasmon decay timescales but have not revealed ejection timescales (*11*). Further, theoretical descriptions of hot-carrier transport are complicated by the multiple distance- and time-scales that must be considered (*12*). Without a comprehensive model, discrepancies between measured charge transport efficiencies cannot be explained (*1*), and the contributions of hot carriers in each of the early decay stages remain unclear (*6*). Debate continues as to whether effects attributed to hot-carrier transport might instead be due to heating (*13*).

Taghinejad *et al.* report a measurement technique for hot-carrier transport dynamics in a nanoscale metal-dielectric transport junction composed of gold and titanium dioxide. The technique excites the metallic nanostructure with pulsed near-infrared light, which generates a current that oscillates on picosecond and subpicosecond timescales (see the figure). The oscillating current then results in terahertz radiation, which is detected with an ultrafast electro-optic detector. Although indirect, this terahertz radiation comes exclusively from hot carriers expelled from the metal, effectively disentangling electron transport from alternative hot-carrier decay pathways with similar timescales.

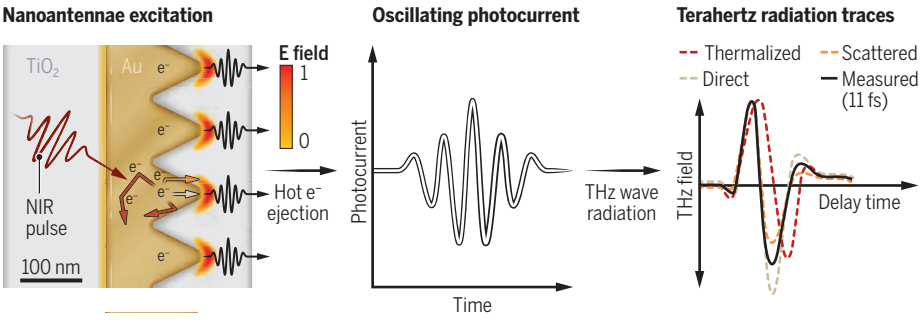
The magnitude of terahertz radiation substantially increased because the authors used asymmetric metal-dielectric junctions to generate the hot carriers. These junctions consisted of thin, gold nanostripes embedded in titanium dioxide, where one side of the stripe was flat, whereas the other had a sawtooth pattern. When excited with near-infrared light, these structures acted as nanoantennae, concentrating electric fields at the tips of the triangular gold teeth. This caused hot carriers to be preferentially created and ejected on the sawtooth side of the stripe. As such, the currents produced at the two nanostripe sides, which progress in opposite directions, were asymmetric and did not counteract each other. With this setup, the photocurrents (and thus the terahertz radiation) were amplified.

The terahertz radiation was modeled as a sum of different contributions to create a comprehensive mechanistic picture. Using this technique, Taghinejad *et al.* differentiated transport of hot carriers at each stage in their energetic decay. They determined that hot-electron ejection dynamics were influenced by both electrons immediately after direct excitation and electrons that had partially decayed through at least one electron-electron collision (scattering). Importantly, however, the terahertz radiation could not be adequately modeled with elevated electron temperatures, suggesting that electrons in thermal equilibrium do not meaningfully contribute to hot-carrier ejection.

By extending the technique reported by Taghinejad *et al.* to metal-dielectric junctions made of other materials and with different sizes, shapes, and environments, transport junction properties can be linked to hot-carrier dynamics. This will lead to predictive design of hot carrier–based sensors, solar cells, and photochemical reactions with energetic structures that capture the maximum number of hot carriers and with materials and dimensions that extend hot-carrier lifetimes. In addition, these measurements provide the opportunity to explore the trajectory of hot-carrier ejection depending on the shapes and orientations of metallic nanostructures (*14*, *15*). This could allow control of the direction or location of hot-carrier transport—for instance, to vary the reaction site or product selectivity within a photocatalytic cell. The approach might also enable the creation of nanojunctions that are sensitive to a variety of light properties such as polarization, orbital angular momentum, or phase, which could allow the development of switchable sensors. Implementation of faster excitation sources may further expand this technique not only to time-resolve hot-carrier transport, but also to serve as an ultrabroadband terahertz source that is useful for imaging and free-space communications. ■

A strategy to follow the mechanism of hot-carrier ejection

Pulsed near-infrared (NIR) light excites an asymmetric gold stripe (nanoantennae) within a metal (Au)–dielectric (TiO_2) transport junction, which results in electric field concentration and asymmetric hot-electron (e^-) ejection (left). This generates a fast-oscillating photocurrent (middle) and a measurable radiated terahertz field that comes exclusively from hot-electron ejection (right; the graph is illustrative). Modeling the terahertz field shows that both direct and scattered electrons contribute to the terahertz trace, but thermalized electrons do not (right).



GRAPHIC: A. FISHER/*SCIENCE*

REFERENCES AND NOTES

1. J. B. Khurgin, *Nanophotonics* **9**, 453 (2019).
2. M. L. Brongersma, N. J. Halas, P. Nordlander, *Nat. Nanotechnol.* **10**, 25 (2015).
3. E. Cortéz *et al.*, *Nat. Commun.* **8**, 14880 (2017).
4. A. O. Govorov, H. H. Richardson, *Nano Today* **2**, 30 (2007).
5. M. Taghinejad *et al.*, *Science* **382**, 299 (2023).
6. P. Christopher, M. Maskovits, *Annu. Rev. Phys. Chem.* **68**, 379 (2017).
7. A. Manjavacas, J. G. Liu, V. Kulkarni, P. Nordlander, *ACS Nano* **8**, 7630 (2014).
8. C. Clavero, *Nat. Photonics* **8**, 95 (2014).
9. H. Reddy *et al.*, *Science* **369**, 423 (2020).
10. L. Du, A. Furube, K. Hara, R. Katoh, M. Tachiya, *J. Photochem. Photobiol. C Photochem. Rev.* **15**, 21 (2013).
11. R. H. M. Groeneveld, R. Sprik, A. Lagendijk, *Phys. Rev. B Condens. Matter* **51**, 11433 (1995).
12. P. Narang, R. Sundararaman, H. A. Atwater, *Nanophotonics* **5**, 96 (2016).
13. B. Y. Zheng *et al.*, *Nat. Commun.* **6**, 7797 (2015).
14. M. Song *et al.*, *Appl. Phys. Rev.* **6**, 041308 (2019).
15. J. Baumberg, J. Aizpurua, M. H. Mikkelsen, D. R. Smith, *Nat. Mater.* **18**, 668 (2019).

ACKNOWLEDGMENTS

The authors are supported by the Air Force Office of Scientific Research (award FA9550-21-1-0312).

10.1126/science.adk6862

MEDICINE

Phage therapy in noncommunicable diseases

Bacteriophages have potential as suppressors of disease-contributing commensal bacteria

By Denise Kviatcovsky[1], Rafael Valdés-Mas[1], Sara Federici[1], Eran Elinav[1,2]

Bacteriophages (phages hereafter) are bacterial viruses that were projected, more than a century ago, to constitute an effective antimicrobial remedy and were widely used as such mainly in the former Soviet Republics. Despite the initial excitement, and given the discovery of antibiotics, interest in antibacterial phage treatment gradually waned. However, exploration of phage treatment in infectious diseases has reemerged to tackle the alarming and globally expanding problem of wide-spectrum antibiotic resistance. Additionally, given the increasing appreciation of the diverse roles of microbiota in modulating health and noncommunicable diseases (NCDs), phages could constitute a promising targeted therapy against commensal bacteria (or pathobionts) that naturally inhabit the body but contribute to NCDs.

Key advantages of phage therapy for infections include phage specificity to their predated bacteria, thereby minimizing damage to the surrounding microbiota; a paucity of eukaryotic host off-target effects; and phage multiplication upon invasion of bacterial cells that enables, upon cellular lysis, propagation and new target engagement. In some cases, phage treatment can even confer antibiotic sensitivity in previously antibiotic-resistant infection (*1*). However, there are also challenges in using phages as anti-infectious agents, including the elicitation of an immune response through their foreign antigenic load and genetic material; the emergence of bacterial antiphage resistance, mediated by spontaneous mutations in bacterial phage receptors that render them resistant to phage engagement; and a multitude of antiphage defense mechanisms. These defenses include restriction-modification endonucleases, CRISPR, and more than 100 other newly discovered defense systems (*2*) that provide bacteria with multiple layers of protection in phage-rich environments, thereby compromising the efficacy of phage treatment.

Microbiotas consist of diverse communities of microorganisms that reside in proximity to mucosal surfaces, where they affect a plethora of host physiological functions. Alterations in microbiota community structure (dysbiosis) are increasingly linked to human diseases, including cardiovascular disease, type 1 and type 2 diabetes, inflammatory bowel disease (IBD), some cancers, and even neurodegeneration. Experimental microbiota-altering approaches, such as personalized nutrition, precision probiotics, microbiota transplantation, and metabolite supplementation, are aimed at modulating the microbiota composition and function toward a healthier configuration. However, targeted suppression of disease-contributing pathobionts without inducing dysbiosis in the surrounding microbial ecosystem remains a daunting unmet need.

As a proof of concept to use phage therapy in NCDs, a phage combination therapy was used to target a clade of newly identified, antibiotic-resistant, human IBD-associated *Klebsiella pneumoniae* (Kp) strains in mice (*3*). This resulted in effective Kp suppression and reduced inflammation and disease severity. Furthermore, the viability and safety of orally administered phages were demonstrated in the gastrointestinal tract of healthy human volunteers. The rational construction of a phage consortium that targeted a bacterial strain or clade of strains of interest, in which each consortium phage member engaged the target strains through a different mechanism, prevented the emergence of phage resistance. Oral administration, when coupled with gastric acid suppression, ensured a viable phage encounter with the gut microbiota, tolerability, and minimal off-target dysbiosis or adverse effects in the participants. Thus, application of the Kp-targeting phage consortia in IBD patients who harbor Kp (which occurs in 38% of the patients assessed in four geographically diverse cohorts) merits further interventional studies.

A virulent phage combination therapy that targets a gastrointestinal Kp clade enriched in patients with primary sclerosing cholangitis (PSC) (*4*) demonstrated a sustained in vitro suppressive effect and an in vivo reduction of fecal Kp, coupled with attenuation of liver inflammation and disease severity in a mouse model of hepatobiliary injury. Similarly, the use of a single virulent phage that targets an alcohol-producing Kp strain alleviated steatohepatitis in mice, without altering the gut microbiota (*5*). Phage intervention was accompanied by suppression of hepatic apoptosis and inflammation pathways, coupled with an enhanced production of the anti-inflammatory cytokine interleukin-10.

Future phage combinations could be applied to couple suppression of disease-contributing commensals with phage-mediated drug delivery into distinct niches. For example, *Fusobacterium nucleatum* has been linked to colorectal cancer (CRC) progression and altered responsiveness to immunotherapy and chemotherapy. A nanosystem, directed by phages engineered to selectively target *F. nucleatum* in the CRC tumor microenvironment in mice, encapsulated the anticancer drug irinotecan in dextran nanoparticles. Such phage-targeted therapy significantly increased the efficacy of the chemotherapy treatment and, at the same time, inhibited the intratumoral growth of *F. nucleatum* (*6*).

In addition to the use of live phages, effector phage proteins may be integrated as an alternative treatment option that enables phage-induced bacterial suppression without the complexities associated with administering live-phage therapeutics, including immunogenicity and resistance. For example, endolysins are peptidoglycan hydrolases that move, through phage-preformed pores, into the murein layer of Firmicutes bacteria, where they degrade the peptidoglycan wall and thereby induce cell death. Using improved synthetic versions of phage lysins could allow treatment of antibiotic-resistant infections (*7*) and potentially NCDs.

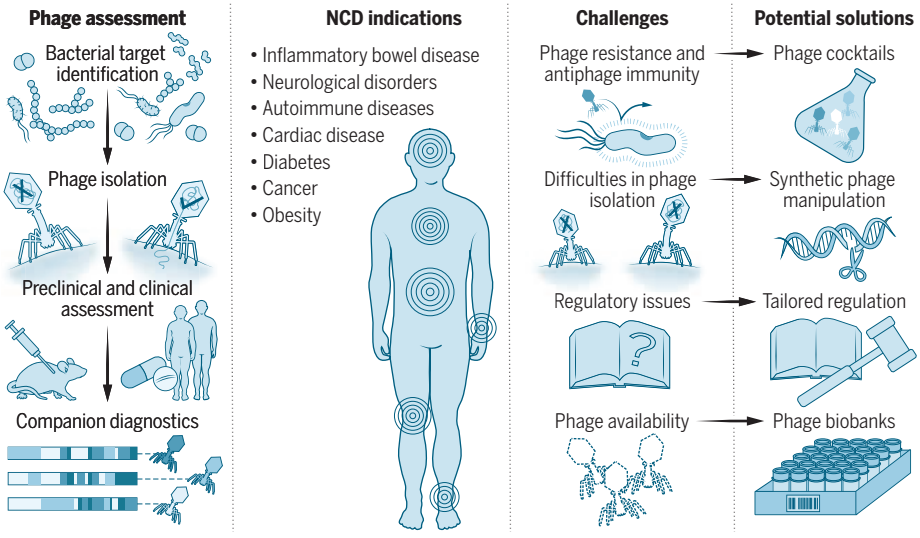
Phage treatment must be tailored to the individual, which may be facilitated by coculture of patient-harvested bacterial targets with potential phages to identify those with antibacterial capacity. As such, metagenomic profiling could be used as a companion diagnostic to identify the dominant disease-contributing pathobionts in a patient, followed by culture-based optimization of phage consortia.

Phage combination therapy for NCDs is at an exploratory phase (see the figure). The challenges of phage therapy for infections, including immunogenicity, resistance, and antiphage bacterial immunity, remain to be explored in the NCD context in prospective long-term studies. Environmental phage isolation, genomic and functional characterization, and testing against bacterial panels are labor-intensive tasks, which are increasingly met with automation, coupled with

[1]Systems Immunology Department, Weizmann Institute of Science, Rehovot, Israel. [2]Microbiome and Cancer Division, DKFZ, Heidelberg, Germany. Email: eran.elinav@weizmann.ac.il

Challenges and prospects of phage therapy

Developing human phage therapy for noncommunicable diseases (NCDs) begins with identifying disease-contributing bacteria and the isolation and testing of potential phages. Complementary phage combinations should prevent the emergence of phage resistance, and these phage cocktails can be tested in preclinical models followed by clinical trials. Companion diagnostics could help identify disease-contributing pathobionts in a patient. However, multiple challenges must be overcome, including difficulties in phage isolation and availability as well as regulatory hurdles.



GRAPHIC: N. BURGESS/*SCIENCE*

the generation of sustainable collections of environmentally and/or clinically isolated phages and phage-encoded proteins. Phage biobanks may provide purified, sequenced, and structurally characterized phages. Such phage repositories may enable an assessment of immunogenicity and other adverse effects that are thoroughly validated by clinical trials (NCT04650607) while facilitating rapid testing of phages and their combinations against their respective bacterial targets.

In cases in which phage isolation proves difficult, several alternative approaches may enable genomic conversion of phage specificity to new targets. In general, temperate phages are abundant but less desirable as therapeutics because they integrate their genome into the bacterial chromosome through a process that involves integrases and repressors. Temperate phages that are rendered integration-defective may still be able to incorporate their genetic material into the bacterium but are less stable and may result, in some cases, in conversion to a virulent state. Such engineered phages may be incorporated into some antibacterial phage combination treatments (*8*). Moreover, the naturally produced bacterial toxin colibactin may awaken dormant prophages into their active forms, enabling their subsequent transformation into virulent, clinically suitable phages (*9*).

The design and selection of clinically usable phages may be facilitated by the application of evolutionary trade-offs. As such, the development of phage resistance may be accompanied by compromised bacterial virulence. This trade-off, experimentally exploited as "phage steering" to modify pathogen virulence and reduce infection severity (*10*), may be likewise explored in NCDs. In cases in which no environmental phages are identified against a pathobiont, synthetic biology may enable the redirection of phage specificity by swapping components of the phage tail (*11*), which specifically recognizes, binds, and injects the genetic material into the bacterial target. Synthetically engineered "phagebodies" may enable further modifications, leading to, for example, the acquisition of a new bacterial target range to allow specificity for resistant bacterial mutants (*12*). Similarly, tail fiber– and CRISPR-Cas–engineered phages significantly improved their efficacy in targeting pathogenic *Escherichia coli* that is associated with fatal infections in mice (*13*).

Phage delivery constitutes a major therapeutic consideration. Orally administered phages face rapidly shifting biophysical conditions in their passage along the human gastrointestinal tract. Variable transit time, periodic food consumption, and a radically altering pH may affect phage viability and activity upon engagement with their bacterial targets. Additional intestinal delivery methods, such as colonoscopy and sustained release and timed-release capsules (*14*), could optimize phage delivery to designated niches along the human gut. Phage engagement with mucins, glycoproteins that form the mucus layer that provides lubrication and protection along the gastrointestinal tract, may reduce phage encounters with mucosal-adherent pathobionts, which often have roles in barrier function and immune tolerance and are disrupted in a variety of NCDs (*15*). Cutaneous phage formulations would optimally enable sufficient penetration into hair follicles, in which many of the skin commensals reside. Administration of phages into the bloodstream may complement local delivery routes and merits future research.

Regulatory challenges must not be overlooked. At present, phage therapy regulation follows, in most countries, an identical route of assessment and approval to that of conventional drug candidates. Nevertheless, because phages are live biological agents, they present numerous differences from traditional medications and so they merit tailored regulatory consideration. For example, approval of combination drug treatments often requires evidence of efficacy for each component. In the case of phages, such an approach would likely fail in many cases, given that bacterial resistance would develop against some or all individual phages when administered as sole interventions. From a safety perspective, the widespread (yet poorly controlled) administration of phages for more than a century, coupled with the lack of known eukaryotic receptors to phages, implies that such treatment should be safe. However, phage-induced innate immune activation and immunogenicity must be closely assessed upon long-term administration to treat NCDs. In balancing these considerations, dedicated supervision would be advisable to tailor the regulatory assessment of phage therapy to its distinct characteristics. Addressing the challenges associated with phage therapy, while exploiting their therapeutic potential in human NCDs, may enable the antibacterial power of phages to be harnessed in human precision medicine. ■

REFERENCES AND NOTES

1. A. Eskenazi *et al.*, *Nat. Commun.* **13**, 302 (2022).
2. S. Doron *et al.*, *Science* **359**, eaar4120 (2018).
3. S. Federici *et al.*, *Cell* **185**, 2879 (2022).
4. M. Ichikawa *et al.*, *Nat. Commun.* **14**, 3261 (2023).
5. L. Gan *et al.*, *Nat. Commun.* **14**, 3215 (2023).
6. D.-W. Zheng *et al.*, *Nat. Biomed. Eng.* **3**, 717 (2019).
7. C. Röhrig *et al.*, *mBio* **11**, e00209-20 (2020).
8. R. M. Dedrick *et al.*, *Nat. Med.* **25**, 730 (2019).
9. J. E. Silpe, J. W. H. Wong, S. V. Owen, M. Baym, E. P. Balskus, *Nature* **603**, 315 (2022).
10. A. R. Burmeister *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11207 (2020).
11. H. Ando *et al.*, *Cell Syst.* **1**, 187 (2015).
12. K. Yehl *et al.*, *Cell* **179**, 459 (2019).
13. Y. E. Gencay *et al.*, *Nat. Biotechnol.* 10.1038/s41587-023-01759-y (2023).
14. D. Shalon *et al.*, *Nature* **617**, 581 (2023).
15. J. J. Barr *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 13675 (2015).

ACKNOWLEDGMENTS

D.K, R. V.-M., and S.F. equally contributed to this Perspective. The authors thank the reviewers and R. Sorek for insightful suggestions. E.E. is a scientific cofounder of DayTwo and BiomX and is an adviser to Aposense, Igen, PurposeBio, and Zoe.

10.1126/science.adh2718

ECOLOGY

Periodical cicadas disrupt forest food webs

Mass cicada emergence satiates birds and makes trees vulnerable to insects

By **John D. Parker**

Every year somewhere in eastern North America, at least one brood of periodical cicadas emerges from underground after spending 13 or 17 years sipping tree root juices. With only weeks of life left, these true bugs (*Hemiptera*) sing loudly to find mates, lay offspring, and then perish by the billions (*1*). Carcasses can stimulate soil nutrient cycling, soil microbes, and tree growth (*2*, *3*), but surprisingly little is known about the real-time effects of living cicadas on food webs. On page 320 of this issue, Getman-Pickering *et al.* (*4*) combine focused experiments with community-science observations before, during, and after an emergence in 2021 to show that mass cicada outbreaks shift the normal diets of entire bird communities, thus indirectly increasing caterpillar abundance and damage to trees. Ultimately, these brief but massive protein windfalls leave a legacy of short- and long-term impacts on food webs far belying their short duration.

> **"...it is rare to see such community-wide shifts in foraging preferences..."**

Getman-Pickering *et al.* analyzed a particularly diverse and broad dataset, a feat made possible with a time-honored tradition in cicada research: community science. Crowdsourcing of data on cicada emergences in the United States started as early as 1851 with Baltimore doctor and silkworm enthusiast Gideon Smith, who enlisted newspaper subscribers to help him delineate 18 "districts" (broods) of cicadas, a tradition that continues to this day with the mobile app "Cicada Safari," which has more than 200,000 users (*5*, *6*). Similarly, Getman-Pickering *et al.* combined their own observations of bird feeding with those from the birding community to document cicada-binging by 82 bird species, including such wildly different species as gnatcatchers (*Polioptila caerulea*), falcons (*Falco* species), turkey vultures (*Cathartes aura*), and trumpeter swans (*Cygnus buccinator*). In just over a month, they documented cicada-feeding by nearly one-fifth of the entire suite of birds known to exist in the study region. It has long been clear that birds eat cicadas (*7*), but it is rare to see such community-wide shifts in foraging preferences, especially among birds more commonly known as either herbivores (for example, trumpeter swans) or scavengers (for example, turkey vultures).

Resource pulses such as insect outbreaks, whale falls to the ocean floor, and acorn masting (production of large numbers of seed) have well documented, substantial effects on the organisms that directly forage these resources (*8*). However, Getman-Pickering *et al.* build on this knowledge by showing that the indirect (and thus often less studied) effects of these resource pulses might be equally important. They conducted tree-based experiments in the years before, during, and after the 2021 emergence to show that bird predation on caterpillar mimics declined by ~75% during the outbreak. At the same time, insect surveys revealed a doubling of the number of herbivorous caterpillars on oak trees and a doubling of insect leaf damage. Thus, the overall picture that emerges is that in a typical year, birds roughly halve oak herbivory by consuming leaf-feeding caterpillars (a classic three-level, top-down trophic cascade), but this respite is lost and the cascade is broken when a cicada emergence shifts the feeding preferences of the bird community. This is an example of apparent facilitation: when one species (cicadas) facilitates another (caterpillars) by altering their interaction strength with a third party (birds). Apparent facilitation has been documented before, but typically among just a few species and rarely at the scale of an entire biome (*9*, *10*).

Pulse events in general, including both insect outbreaks and climatic extremes such as heat waves, are helpful teaching moments because they can reveal ecosystem response to chronic stressors. The loss of bird predation and concomitant rise in herbivory observed by Getman-Pickering *et al.*, for example, is functionally analogous to the anthropogenically driven trophic skew that has occurred over the past few centuries, in which the decimation of apex predators such as wolves led to the rise of middle-ranked consumers, with subsequent negative effects lower down the trophic chain (*11*).

In the case of the 2021 cicada emergence, the top-down services provided by birds were regained after just 1 year, providing optimism that ecosystems are resilient and functions can be regained when top-down trophic forces are restored, which is indeed sometimes the case (*12*). Questions remain, however, about the long-term effects of pulse events on food webs. Previous work has provided an intriguing answer for cicadas (*13*), showing that although damage caused by egg-laying initially suppresses oak acorn production, in later years acorn crops are elevated, potentially from nutrient pulses caused by cicada carcasses.

There is a critical need to detect the reorganization of food webs in response to climate change and other stressors, but as shown by Getman-Pickering *et al.*, monitoring even just a few trophic levels is often too daunting for any one research group to accomplish alone. Moreover, because they rapidly track changes in the environment, mobile organisms such as birds can be early-warning indicators for global change (*14*), yet they are typically undersampled even in fairly large biodiversity databases (*15*). However, ecology has been progressively moving toward community science–assisted biodiversity-monitoring networks and collective experiments distributed over large geographical ranges, resulting in expanded geographic and taxonomic resolution. Moving forward, linking these existing biodiversity-monitoring networks into global networks (*15*) will hopefully provide real-time information on changes in biodiversity and provide a pathway for turning data into meaningful conservation action. ■

Smithsonian Environmental Research Center, Edgewater, MD, USA. Email: parkerj@si.edu

REFERENCES AND NOTES

1. C. Simon *et al.*, *Annu. Rev. Entomol.* **67**, 457 (2022).
2. L. H. Yang, R. Karban, *Ecology* **100**, e02705 (2019).
3. H. Setälä *et al.*, *Ecology* **103**, e3773 (2022).
4. Z. L. Getman-Pickering *et al.*, *Science* **382**, 320 (2023).
5. G. Kritsky, *Am. Entomol.* **67**, 40 (2021).
6. I. Graber-Stiehl, *Science* **372**, 1022 (2021).
7. W. D. Koenig, A. M. Liebhold, *Ecology* **86**, 1873 (2005).
8. L. H. Yang *et al.*, *Ecol. Monogr.* **80**, 125 (2010).
9. J. M. Levine, *Ecology* **80**, 1762 (1999).
10. D. E. Lee *et al.*, *Ecol. Evol.* **6**, 8402 (2016).
11. L. R. Prugh *et al.*, *BioScience* **59**, 779 (2009).
12. J. A. Guyton *et al.*, *Nat. Ecol. Evol.* **4**, 712 (2020).
13. W. D. Koenig *et al.*, *Am. Nat.* **201**, 755 (2023).
14. T. J. Bartley *et al.*, *Nat. Ecol. Evol.* **3**, 345 (2019).
15. A. Gonzalez *et al.*, *Nat. Ecol. Evol.* 10.1038/s41559-023-02171-0 (2023).

10.1126/science.adk5880

RETROSPECTIVE

W. Jason Morgan (1935–2023)

Discoverer of plate tectonics and mantle plumes

By **Richard Hey**[1] *and* **Jason Phipps Morgan**[2]

William Jason Morgan, the geophysicist who transformed Earth sciences, died on 31 July. He was 87. Morgan discovered plate tectonics, a concept that influences scientists' understanding of almost every branch of geology.

Born on 10 October 1935 in Savannah, Georgia, Morgan earned a BS in physics at the Georgia Institute of Technology in 1957 and then served 2 years in the navy, where he taught nuclear physics and navigation to submariners. In 1959, he married Cary (née Goldschmidt), and in 1964 he earned a PhD in physics from Princeton University. Because his dissertation involved geophysics, his adviser, Bob Dicke, invited the geology department chair, Harry Hess, to be on his thesis committee. Hess offered Morgan a postdoc position and, a year later, hired geophysicist Fred Vine. These colleagues drew Morgan into seafloor spreading "with a bang," as he described it. After 4 decades, Morgan retired from Princeton in 2003 and—to be near his daughter, Michèle—moved to Harvard University to work as an emeritus professor.

In the 1960s, the field of geology was changing from a "fixist" view of Earth, which posited that the continents and ocean basins had always existed in their present positions, to a "mobilist" view, which recognized large horizontal land movements. Hess and Vine were at the forefront of this transition. Continental drift had been hypothesized 50 years earlier but had since been rejected. In 1960, Hess proposed that seafloor spreading could resolve some of the objections to continental drift, and Vine showed that seafloor spreading combined with episodic reversals of Earth's magnetic field could explain the puzzling seafloor magnetic stripes. In 1966, the transform fault hypothesis, which was contingent on seafloor spreading, was corroborated seismically, and Vine correlated the symmetric Eltanin-19 magnetic anomalies and others from the global mid-ocean ridge system with an improved magnetic reversal timescale. By year's end, most marine geophysicists agreed that seafloor spreading was real.

[1]School of Ocean and Earth Science and Technology, University of Hawai'i at Mānoa, Honolulu, HI, USA.
[2]Department of Ocean Science and Engineering, Southern University of Science and Technology, Shenzhen, China.
Email: hey@soest.hawaii.edu; japhmo@icloud.com

In January 1967, just days after accepting a position as assistant professor, Jason read a paper showing that the Great Pacific Fracture Zones were almost great circles. He realized from his navy experience with map projections that they were instead small circles concentric about a pole of rotation. Dropping everything to pursue the implications, he worked out the principles of plate tectonics within weeks. J.P.M., Morgan's son, recalls his mom saying that Morgan worked nonstop during this time, making her nervous about what life was going to be like married to an academic.

PHOTO: DENISE APPLEWHITE/PRINCETON UNIVERSITY DEPARTMENT OF GEOSCIENCES

In April 1967, Morgan presented the quantitative theory of plate tectonics, which subsumed and explained seafloor spreading and continental drift. This work extended J. Tuzo Wilson's transform fault work to the sphere. It assumed that Earth's plates were rigid, giving the theory mathematical rigor. Continents drifted because they were embedded in moving tectonic plates. Seafloor spreading occurred where plates moved away from each other. There were three kinds of plate boundaries because plates could move away from (ridges), move toward (trenches), or slide past each other (transform faults). He showed how to use spherical geometry along with poles of rotation and triple junction analysis to quantitatively predict plate motions, even in the absence of data for a particular plate pair. Tectonics was now a quantitative science, and the causes of mountains, earthquakes, and volcanoes became clear.

A few years later, Morgan resurrected Wilson's hotspot hypothesis, proposing that island chains extend away from relati fixed melting anomalies—which at the was not generally accepted—and that these melting anomalies form at the top of deep mantle plumes, narrow columns of hot rock rising from the deep mantle that form the upwelling counterflow to downgoing tectonic plates. Morgan conceived of this groundbreaking idea after hearing two talks at a 1970 workshop, one on the existence of large geoid anomalies in Hawai'i and Iceland and the other on how the volcanic rock types in Hawai'i and Iceland differ from those of mid-ocean ridges. Morgan suddenly realized that if deep mantle upwellings underlay these volcanic hotspots, then both the geophysical and geochemical observations could be explained. At the time, R.H. was working on Galápagos tectonics for his PhD with Morgan, and several groups had presented seemingly fatal geometric objections to the hotspot interpretation in the region. Morgan told R.H. that he was sure that Galápagos was a hotspot and urged him to figure out where the other groups had gone wrong. Galápagos is now considered to be a type-example hotspot.

Morgan once told R.H. that a perfect adviser secures funding and then turns students loose on problems. Yet when his own students got stuck, he would usually point the way out. His insight about Earth was remarkable, especially considering that, as a physics student, he never took a geology class. He was a gifted programmer, an effective and caring teacher, and a supportive mentor. During a sabbatical in Hawai'i—a trip that deepened the local understanding of hotspots and started J.P.M. on his own path to a career in geophysics—the family and R.H. backpacked to the Mauna Loa summit caldera, which a geologist friend had assured them would be "a piece of cake." With his characteristic sense of humor, Morgan used that phrase ever after to refer to any extremely hard problem. Although the million-dollar view from the caldera's summit cabin outhouse was worth the effort, J.P.M. also remembers his surprise upon opening his heavy pack to find beer, which Morgan had surreptitiously stashed there to celebrate the summit ascent.

Morgan was honored with many awards for his achievements, among them the Walter H. Bucher Medal in 1972, the Japan Prize in 1990, and the National Medal of Science in 2002. An understanding of plate tectonics was nothing short of a scientific revolution, yet Morgan always kept his warmth and humility. His scientific career gave him a path for creative satisfaction and friendships around the world. He treasured these far more than any awards. ■

10.1126/science.adk9812

POLICY FORUM

ECONOMICS

Building an alliance to map global supply networks

New firm-level data can inform policy-making

By **Anton Pichler[1,2], Christian Diem[1,2], Alexandra Brintrup[3], François Lafond[4], Glenn Magerman[5], Gert Buiten[6], Thomas Y. Choi[7], Vasco M. Carvalho[8,9,10], J. Doyne Farmer[4,11], Stefan Thurner[1,11,12,13]**

The global economy consists of more than 300 million firms, connected through an estimated 13 billion supply links [see supplementary materials (SM)], that produce most goods and services. It has long been unthinkable to analyze the world economy at the firm level, even less so its intricate network of supply chain linkages. This blind spot has left us ill-prepared to make fast and well-informed decisions, begetting, for example, prolonged shortages in raw materials and critical medical supplies during the COVID-19 pandemic. Now, the availability of new data and recent methodological advances allow us to reconstruct a large share of the global firm-level supply network. Because mapping this network is likely to continue to improve, it is essential to initiate a discussion about responsible management and effective use of these data for the global public good. This requires new collaborative efforts between nations, their public institutions, international organizations, the private sector, and scientists.

Potential applications of a global supply chain map include monitoring and improved management of the green transition, reducing tax evasion and corruption, strengthening human rights through supply chain transparency, identifying and monitoring systemic risks and systemically important firms, and the design of globally secure basic provisioning systems for food and medication. Toward such ends, research has contributed to a better understanding of the functioning of supply chains yet has been constrained by data limitations. Although highly granular data are available for single "focal" firms where the direct suppliers and customers are known, such data are not connected to the rest of the economy, and no network perspective is possible. By contrast, when looking at how supply chains affect economy-wide phenomena such as gross domestic product (GDP), business cycles, or inflation, one is typically constrained to highly aggregated data in the form of input-output relationships between a few dozen industrial sectors.

Until now, such data limitations have made it impossible to integrate granular product- and firm-level expertise with macro-level, economy-wide perspectives. Recently, high-resolution maps of firm-level supply networks have been charted for individual economies (*1*). However, because they are not connected, the global image—the one that really counts in a globalized economy—remains fragmented. Even national maps are typically unavailable to researchers and policy-makers, which limits their potential usefulness for addressing societal problems.

A NEED FOR GLOBAL NETWORK DATA

In 2021, supply disruptions caused an estimated 2% loss of global GDP (roughly USD 1.9 trillion) and substantially contributed to high inflation (*2*). In a globalized economy, local shocks to individual firms can spread to geographically distant firms through several tiers in the supply chain. Spotting such spreading dynamics ahead of time is impossible with aggregate, industry-level data or granular but incomplete supply chain data.

The potential for modeling the propagation of economic shocks in a detailed manner has been shown for individual countries for which large-scale firm-level supply network data are available. For example, in 2011, the Great East Japan Earthquake affected only four prefectures directly, but the economic shocks propagated throughout the economy. Based on detailed Japanese supply network data of several million firm-level supply links, it was possible to model the indirect country-wide economic impacts in much detail (*3*).

A new generation of economic models becomes possible when calibrated to granular supply network data, which enables us to address "what-if" scenarios for direct use in policy interventions at the micro- and macro-level. For example, using the detailed Japanese dataset, researchers predicted a recession of 2.2% with a probability exceeding 10% if a Tōkai earthquake with magnitude greater than 8 occurs, which is considered likely in the coming decades (*3*).

In some countries, value-added tax (VAT) data are collected on the level of individual trades, which allows the recovery of virtually all domestic business-to-business trades. These data have been used to quantify the economic importance of individual firms by modeling their direct and indirect macroeconomic impact in case of failure (*4*). Analogously to financial networks, we could leverage detailed supply network data to monitor the systemic risk levels posed by individual firms and incorporate this into economic regulation.

Complex supply relationships are also highly susceptible to tax fraud. Between 2016 and 2019, the annual VAT-related tax gap has been estimated to EUR 120 billion within the European Union (EU). "Carousel fraud" alone, which involves the intentional circulation of goods among companies and countries to avoid VAT payments, accounts for around EUR 50 billion in lost tax revenues annually. The ability to measure individual supply linkages between firms could be immensely valuable in detecting and preventing such fraudulent activities.

As policy-makers increasingly recognize, supply chains are critical for ensuring compliance with tax laws, human rights, and environmental standards. Recent policy initiatives such as the US Uyghur Forced Labor Prevention Act or the EU Supply Chain Act aim to hold companies accountable for monitoring and upholding social and environmental standards throughout their supply chain.

Managing the transition to a carbon-neutral economy requires a detailed understanding of supply chain relationships. Because of the lack of data, it is, at present, impossible to accurately assess and monitor the carbon intensity of products and indirect emissions caused by individual firms. Deploying new green technologies on a large (global) scale while phasing out fossil technologies will substantially "rewire" the supply network, resulting in heterogeneous geographical impacts on employment, energy prices, and financial stability. These changes will create new geopolitical dependencies and shift vulnerabilities in economic production by introducing new critical materials.

[1]Complexity Science Hub, Vienna, Austria. [2]Vienna University of Economics and Business, Vienna, Austria. [3]Institute for Manufacturing, University of Cambridge, Cambridge, UK. [4]Institute for New Economic Thinking and Mathematical Institute, University of Oxford, Oxford, UK. [5]European Center for Advanced Research in Economics and Statistics, Université Libre de Bruxelles, Brussels, Belgium. [6]Centraal Bureau voor de Statistiek, The Hague/Heerlen, Netherlands. [7]Department of Supply Chain Management, W. P. Carey School of Business, Arizona State University, Tempe, AZ, USA. [8]Faculty of Economics, University of Cambridge, Cambridge, UK. [9]Centre for Economic Policy Research, London, UK. [10]Alan Turing Institute, London, UK. [11]Santa Fe Institute, Santa Fe, NM, USA. [12]Section for Complex Systems, Medical University of Vienna, Vienna, Austria. [13]Supply Chain Intelligence Institute Austria, Vienna, Austria. Email: pichler@csh.ac.at

A NEW ERA OF MICRODATA

Until very recently, supply network data has been limited to a few hundred firms or industries (traditional input-output tables). Only in the past decade has a supply network data revolution started with several independent lines of progress, including payment data, VAT data, and various reconstruction methods. Through these efforts, the scope of supply chain data could be increased by several orders of magnitude (see the figure).

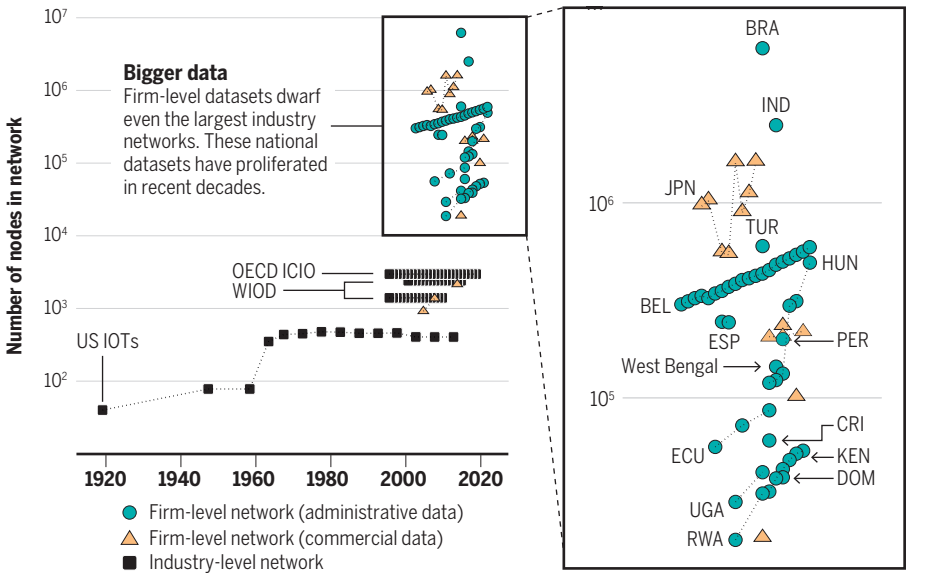
The ongoing expansion of supply chain data is also driven by the private sector. As demonstrated by recent investments in supply chain analytics firms worth hundreds of millions USD, a better understanding of supply dependencies is highly valued information. Datasets collected by firms that provide business intelligence services can be extensive, including more than 200,000 firms with more than 500,000 supply linkages (*1*), and have proven vastly useful for reconstructing supply chain relationships. These datasets are derived from a wide range of sources, including credit-rating agencies, firm disclosures on public filings, business reports, capital market presentations, and press releases. Limitations of commercial datasets include a bias toward large publicly listed firms, uncertain quality assessment, and the lack of methodological transparency (*5*).

To overcome some of these issues, payment data and bank transaction data have recently been tapped to reconstruct supply linkages between firms (*6*). However, these data are typically not readily available for research or supporting policy-making, and it can be challenging to reliably extract supply dependencies from payment data. The most-complete information on firm-level supply networks comes from administrative VAT records. Out of the 170 countries that collect VAT, researchers have mapped virtually all domestic trades between firms for Belgium, Chile, Kenya, Turkey, Ecuador, Costa Rica, Uganda, Hungary, Spain, Rwanda, multiple states of India, and the Dominican Republic, and similar efforts are ongoing for more countries. These datasets represent an unprecedented window into the microstructure of buyer-customer relationships that enable us, for the first time, to characterize entire national economies as complex supply networks (*7*, *8*).

These datasets show that individual firms can have tens of thousands of suppliers, which themselves are connected to a vast network of supply linkages. Because firms can only observe their direct suppliers and customers, it is very difficult for them to reconstruct the upper tiers of their supply chains. Instead, using already collected tax data could provide an extensive picture without additional administrative burdens for companies.

The rise of large-scale firm-level supply network data

In firm (industry)–level datasets, nodes represent firms (economic sectors or sector-country pairs). Administrative data are derived from value-added tax (except in the case of Brazil). US input-output tables (IOTs), which are among the most detailed in the world, illustrate national industry-level data. WIOD and OECD ICIO illustrate intercountry IOTs that are primarily derived from national IOTs and trade data.



BEL, Belgium; BRA, Brazil; CRI, Costa Rica; DOM, Dominican Republic; ECU, Ecuador; ESP, Spain; HUN, Hungary; IND, India (data only for five states); JPN, Japan; KEN, Kenya; OECD ICIO, Organisation for Economic Co-operation and Development Inter-Country Input-Output tables; PER, Peru; RWA, Rwanda; TUR, Turkey; UGA, Uganda; WIOD, World Input-Output Database.

A MAPPING BLUEPRINT

Even though VAT data enable us to reconstruct complete domestic supply networks at the firm level, it is necessary for many, if not most, essential applications to combine these for different countries. Securing the provision of critical goods, monitoring human rights, fighting cross-border tax fraud, and monitoring the carbon footprint of individual products will remain ineffective without a granular, comprehensive, and international map of supply linkages.

Extending the detailed domestic supply networks across country borders requires implementing regionally and internationally coordinated strategies at the same time. First, it is necessary to collect the domestic firm-to-firm trades through granular VAT records, for example, using e-invoicing. Such systems can reduce administrative burdens borne by firms and substantially contribute to tax compliance (*9*). Thus, treasuries and tax offices should have strong incentives to do this. Several countries have successfully implemented efficient reporting standards that can serve as role models for others (see SM). Among those countries, there are several small and developing economies with limited resources, demonstrating the technical and economic feasibility of collecting firm-level supply network data through VAT records.

The next step requires connecting different countries on the basis of trade data. In the EU, the firm-level trade of goods between member states is collected by the individual countries and reported to the EU Intrastat system. If the EU were to extend these data to all goods and services and merge them with the domestic VAT data, this could result in the first comprehensive multicountry firm-level supply network, representing nearly 20% of world GDP. The recent proposal VAT in the Digital Age (ViDA) by the European Commission could provide the legal framework and set a new international standard.

The EU could further enlarge its map of supply dependencies to partially include nonmember countries by merging its supply network data with customs data. Customs data feature excellent coverage and detail on international firm-level trade and have already been linked with domestic VAT-based supply networks (*10*). The supply network of EU firms and their linkages to non-EU trading partners would yield a granular view into supply chain dependencies that cover almost 40% of global trade.

A strategy to reconstruct the EU firm-level supply network could provide a blueprint for other economic and trade communities like Mercosur, the East African Union, or the Association of Southeast Asian Nations. Various Latin American, African, and Asian countries, including China and several states of India, already have the relevant data-collection procedures in place.

GRAPHIC: M. HERSHER/*SCIENCE*

One of the major countries that does not collect VAT is the United States. However, the United States is fairly well covered in most commercial datasets and has extensive commodity flow surveys. The United States is actively developing supply chain maps and considering the creation of a centralized supply chain data hub.

Once supply networks for different economic blocs are mapped, the next step would be to connect these through customs and trade data. To achieve a coherent database of international supply linkages, harmonized standards on data collection and formatting should be developed, for example, by building on best practices provided by individual countries. As happened for the development of national accounts, such standardization efforts could be coordinated by international organizations such as the United Nations, the International Monetary Fund (IMF), the World Bank, or the Organisation for Economic Co-operation and Development (OECD). A big advantage of internationally harmonized economic data, such as national accounts, is that they can be matched with other relevant datasets, enabling analysis of the interaction of sectoral input-output dependencies with financial and environmental accounts or occupational statistics. A coordinated international effort could establish such a standard at the firm level.

However, because detailed supply chain data are highly sensitive, they will not be shared between regions with strained economic and political relationships in the foreseeable future. Thus, it is critical to develop strategies for filling in the blind spots. Commercial datasets paired with network reconstruction methods can be used for reconstructing critical international supply chain dependencies. In recent years, much progress has been made in reconstructing missing information in complex networks, but these methods need to be carefully adapted to the context of supply connections.

Reconstruction methods for supply networks involve a mix of economic theory, careful accounting, combining public and commercial datasets, and state-of-the-art statistical methods. For example, we know a firm is more likely to buy from another firm if they are geographically close and have complementary product portfolios. If we observe enough interesting features of firms, we can train machine-learning algorithms to recognize pairs of firms that are likely to trade. Other approaches have built on natural language processing, knowledge graphs, maximum entropy models, and leveraging telecommunication data [e.g., (*11*–*13*)]. As in the past, these methods and new ones will evolve as more data become available.

A STRONG ALLIANCE

Drawing a trusted and comprehensive picture of international supply linkages that can effectively be used for policy-making requires integrating multiple datasets, developing analytical tools, and establishing secure infrastructure for storing and processing sensitive information. Advancing this agenda requires a strong international alliance of various stakeholders, including national governments, statistical institutes, international organizations, central banks, the private sector, and the scientific community.

Supply chain data can be weaponized if they get into the wrong hands (*14*), necessitating the highest standards for securely storing, sharing, and regulating the access and use of the data. Statistical offices have already developed protocols for providing controlled access while maintaining strict data security and privacy protections for national supply networks and other types of sensitive data, for example, individual-level data in microdata centers. One example where highly sensitive data is collected and provided for research and policy support in a trustworthy manner is the European Health Data Space. Similar approaches should be adopted for international supply chain data to ensure the highest data security and privacy standards while enabling meaningful use for research and policy.

As primary users of these data, national governments would benefit from and must play a key role in data collection and administration. The lack of granular economic microdata frequently hinders targeted policy approaches, resulting in ineffective redistribution of public funds. Transparent and safe access to detailed supply network data will enable governments to design evidence-based policies with greater precision. For example, such data could be used for identifying indirect exposures to natural disasters or upstream supply chain bottlenecks for critical goods, such as medical supplies. In this case, governments could take resilience-enhancing measures such as incentivizing companies to diversify their suppliers.

Central banks could similarly benefit, because extensive firm-level supply network data will considerably improve our understanding of inflation dynamics and how the propagation of economic shocks affects financial stability [e.g., (*15*)]. Central banks are already at the forefront of working with and making these data accessible for scientific economic analysis. As data availability and methodological development progress, they could integrate this knowledge into monetary policy and financial market supervision.

At present, there are various independent efforts at both the national and supranational levels to chart specific parts of the global supply network. Notable examples include the EU proposal for a directive on corporate sustainability due diligence, the US Supply Chain Disruptions Task Force, and the UK Department of International Trade supply chains resilience framework. However, without a concerted approach, these maps will remain fragmented and inadequate for addressing critical societal challenges. International organizations, including the IMF, the World Bank, and the OECD, have considerable expertise in harmonizing international datasets and should play key roles in scaling up these efforts to the international level. Building this alliance would result in a comprehensive map of international firm-level supply connections, a foundation for economic analyses and policies on both the national and international levels. ■

REFERENCES AND NOTES

1. A. Bacilieri, A. Borsos, P. Astudillo-Estevez, F. Lafond, "Firm-level production networks: What do we (really) know?" Working paper 2023-08, Institute for New Economic Thinking, University of Oxford, Oxford, May 2023.
2. O. Celasun, N.-J. H. Hansen, A. Mineshima, M. Spector, J. Zhou, "Supply bottlenecks: Where, why, how much, and what next?" Working Paper WP/22/31, International Monetary Fund, 17 February 2022.
3. V. M. Carvalho, M. Nirei, Y. U. Saito, A. Tahbaz-Salehi, *Q. J. Econ.* **136**, 1255 (2020).
4. C. Diem, A. Borsos, T. Reisch, J. Kertész, S. Thurner, *Sci. Rep.* **12**, 7719 (2022).
5. G. Culot, M. Podrecca, G. Nassimbeni, G. Orzes, M. Sartor, *J. Supply Chain Manage.* **59**, 3 (2023).
6. L. N. Ialongo *et al*., *Sci. Rep.* **12**, 11847 (2022).
7. T. Y. Choi, K. J. Dooley, M. Rungtusanatham, *J. Op. Manage.* **19**, 351 (2001).
8. E. Dhyne, G. Magerman, S. Rubínová, "The Belgian production network 2002–2012," Working paper 288, National Bank of Belgium, Brussels, October 2015.
9. D. Pomeranz, *Am. Econ. Rev.* **105**, 2539 (2015).
10. E. Dhyne, A. K. Kikkawa, M. Mogstad, F. Tintelnot, *Rev. Econ. Stud.* **88**, 643 (2021).
11. A. Brintrup *et al*., *Complexity* **2018**, 9104387 (2018).
12. E. E. Kosasih, A. Brintrup, *Int. J. Prod. Res.* **60**, 17 (2021).
13. L. Mungo, F. Lafond, P. Astudillo-Estévez, J. D. Farmer, *Int. J. Dyn. Control* **148**, 104607 (2023).
14. H. Farrell, A. L. Newman, *Nature* **605**, 219 (2022).
15. C. Duprez, G. Magerman, "Price updating in production networks," Working paper 352, National Bank of Belgium, Brussels, October 2018.

ACKNOWLEDGMENTS

The authors thank L. Yang for the visualization, J. Yang for help with the data, and the participants at the *First Interdisciplinary Workshop on Firm-Level Supply Networks* in November 2022 at the Complexity Science Hub, Vienna, for stimulating discussions. A.P. received funding from the James S. McDonnell Foundation (2020-1640). C.D. received funding from the OeNB Anniversary Fund (P18696) and the Austrian Research Promotion Agency (FFG) (886360). F.L. received funding from Baillie Gifford and from UK Research and Innovation through Economic and Social Research Council grant PRINZ (ES/W010356/1). G.M. received funding from Actions de Recherche Concertées (ARC) grant (20204850) and Fonds National de la Recherche Scientifique (FNRS) (H.P046.20). V.M.C. received funding from European Research Council Consolidator Grant MICRO2MACRO (#GAP-101001221). J.D.F. received funding from Baillie Gifford. S.T. received funding from FFG (882184 and 886360).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adi7521

10.1126/science.adi7521

Lia Halloran's vivid paintings capture the drama and mystery of various cosmological events.

BOOKS *et al.*

ASTRONOMY

Fantastical phenomena, illustrated

An astrophysicist and an artist craft evocative portraits of warping spacetime

By **David Kaiser**

Not so long ago, black holes seemed enigmatic, even to the experts who studied them. The idea was both alluring and fantastical: hypothetical hunks of matter compressed to enormous densities, warping their surrounding spacetime so strongly that even light could not escape. Although the phenomenon seemed to follow from the equations of Albert Einstein's general theory of relativity, empirical evidence that black holes actually existed remained suggestive but indirect until February 2016, when members of the Laser Interferometer Gravitational-Wave Observatory (LIGO) published data from the first compelling detection of gravitational waves. Since then, the LIGO Scientific Collaboration has published dozens of comparable signals, most of them attributable to black hole collisions.

In April 2019, another globe-spanning international collaboration, the Event Horizon Telescope, released its own shocking first detection: a now-iconic image of the black hole at the center of the nearby galaxy Messier 87. The black hole sits amid a swirl of superheated plasma, a shadow warping the light from nearby matter in exactly the Escher-like, looping fashion that theoretical physicists had predicted decades earlier, on the basis of Einstein's equations. In May 2022, the group released an equally stunning image of the black hole anchoring our own Milky Way galaxy.

More recent data, released in late June 2023 by the North American Nanohertz Observatory for Gravitational Waves, or NANOGrav, identify a constant thrum of cosmic static: a stochastic background of gravitational waves, rattling around the cosmos from every direction in the sky. The NANOGrav data are most readily accounted for by the motions of a huge population of supermassive black holes, dotted across the Universe, each object millions or billions of times more massive than our Sun.

The Warped Side of Our Universe
Kip Thorne and Lia Halloran
Liveright, 2023. 240 pp.

These days, black holes seem not only possible but plentiful. Yet being ubiquitous has not made them mundane. In *The Warped Side of Our Universe*, physicist Kip Thorne and artist Lia Halloran blend evocative imagery with prose-poem narration to share some of the wilder implications of exotic gravitational phenomena with nonspecialist readers.

Thorne has been a pioneer in the study of gravitation for more than half a century; he shared the Nobel Prize in Physics in 2017 as part of the LIGO leadership team. He also has decades of experience describing esoteric phenomena in accessible ways. Since 2009, Thorne has struck up a series of collaborations with artists, filmmakers, and musicians, experimenting with novel ways to convey the majesty and the strangeness of black holes, gravitational waves, and more-fantastical—as yet hypothetical—phenomena such as wormholes and time travel. Halloran, an accomplished visual artist whose work has been featured in two dozen solo exhibitions around the world, developed an interest in depicting scientific topics during an early stint at the famous Exploratorium in San Francisco and studied astronomy while pursuing her graduate training in painting.

Thorne and Halloran worked on this book over the course of 13 years, a period that included the stunning first results from LIGO and the Event Horizon Telescope and that set the stage for NANOGrav. The authors' process, as recounted in a brief preface, was every bit as looping and nonlinear as Einstein's equations—informal discussions led to sketches and short drafts of prose; refined drawings inspired updated wording, and vice versa. Midway through their efforts, Thorne decided to transform his sparse written contributions into verse, to better complement the swirling, ocean-like eddies emerging in Halloran's paintings.

The result is an ambitious art-science artifact inviting readers to ponder black holes, wormhole-like tunnels through spacetime itself, and the gravitational-wave signals emanating from such strange cosmic constituents. The material is grounded in cutting-edge research. In some places, Thorne draws on novel vocabulary that he and his graduate students coined only a few years ago to help make sense of their sophisticated numerical simulations, while Halloran's engrossing images convey a visceral sense of stormy spacetime ruptures. My favorite section features Thorne's personal reflections on the LIGO collaboration, which involved human-scale interactions at least as complicated as anything described by Einstein's mathematics.

The main body of the book aims more to evoke than to explain. Thorne provides helpful resources in the back of the book, with a more traditional popular-science exposition, plus a glossary, timeline, and annotated bibliography of additional materials to explore. Complementing these other excellent resources, *The Warped Side of Our Universe* offers a creative, original portrait of a scientific community, and a broader universe, in motion. ■

The reviewer is at the Program in Science, Technology, and Society and the Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Email: dikaiser@mit.edu

10.1126/science.adk4698

ILLUSTRATION: LIA HALLORAN

HISTORY OF SCIENCE

Co-creating maps and atlases

Cooperation was critical to the creation of two major scientific resources

Rivals: How Scientists Learned to Cooperate
Lorraine Daston
Columbia Global Reports, 2023. 160 pp.

By **Caroline S. Wagner**

In her new book, *Rivals: How Scientists Learned to Cooperate*, historian of science Lorraine Daston interrogates two massive scientific undertakings: the Carte du Ciel, a 19th-century effort to map the positions of millions of stars, and the post–World War II revision to the International Cloud Atlas, a resource for meteorologists that sought to standardize vocabulary used to describe clouds. Why, she wonders, "did the people and institutions charged with completing the star map and updating the cloud atlas decide to cooperate rather than defect?" They were rivals, after all, were they not?

Daston dubs the Carte du Ciel the "diplomatic model" of rivalrous cooperation. It began in Paris in 1887, gathering observers from 16 countries to map the stars. Annual meetings brought together scientists from 18 observatories to coordinate measurements and share mapping tasks. The international project persevered, and grew, for 83 years, with contributions from many nations.

The "voluntarist model," as represented by the International Cloud Atlas project, inaugurated in Munich in 1891 with delegates from many countries, is the second model detailed by Daston. The eventual atlas created in 1896—*Atlas international des nuages*—emerged from years of lettered communications and conferences representing, according to Daston, "an international consensus of their colleagues."

Although the question is posed, she does not explore why delegates and volunteers did not "defect" from these projects. Instead, she puts forth another question: Were the collaborations' communications regulated by "an invisible hand"?

The reviewer is at the John Glenn College of Public Affairs, Ohio State University, Columbus, OH 43210, USA.
Email: wagner.911@osu.edu

The very first experimental philosophers, in forming the Royal Society of London in 1660, vowed in the group's motto to "take nobody's word for it" (*Nullius in verba*)—no ancient sage, no holy book—without proof of veracity. Evidence necessitated documentation.

Daston reports that a "Republic of Letters" thus grew up among early modern scientists as they documented and shared the results of research and observation.

Consensus was key to the successful post–World War II revision of a cloud reference manual.

This robust correspondence eventually led to the development and growth of scholarly journals.

Describing the value of documentation, in 1665, Robert Hooke wrote in his groundbreaking book *Micrographia*: "And though there have been many curious men before us…their observations [were] so scattered, and they have had no means of communicating the same to one another, and of joining their observations together, that what one man hath done hath been almost lost to another; whereas now, by the help of writing, and printing, and by the institution of this Society, all our observations may be communicated together, and we may all build one upon another's foundations" (*2*). Does this sound like rivalry?

This incipient communications system originated from an "Invisible College" of self-organizing didacts (*1*), with features similar to Adam Smith's invisible economic hand. The Invisible College, or Republic of Letters, brought together people who, liberated from scholasticism, used novel techniques to accumulate information. The growing network amassed evidence for justified belief—fact heaped on fact, validated by debate—which Daston misinterprets as rivalry among "warring savants."

Self-organizing groups, operating as a communications network, respond and adapt to their surroundings in much the same way as markets. This fact calls into question Daston's dismissal of the state as being "more hindrance than help to the Republic of Letters." It is not, in my view, a coincidence that the scientific revolution appeared alongside the state in the mid-17th century. The state structure protected ideas, financed research, and procured results, the very conditions that allowed the Republic of Letters to rise. Indeed, the Royal Navy gave the Royal Society its first commission: Help us manage the forests needed to build warships.

A cloud atlas captures fleeting shapes where the human eye will imagine dinosaurs and dragons. In the clouds of history, Daston sees shapes of warring rivals and egos vying for recognition. Yet cooperation, rather than rivalry, marks the network of the Invisible College—the feature has long been what makes science science. ■

REFERENCES AND NOTES

1. D. Crane, *Invisible Colleges: Diffusion of Knowledge in Scientific Communities* (Univ. of Chicago Press, 1972).
2. R. Hooke, *Micrographia: or some physiological descriptions of minute bodies made by magnifying glasses. With observations and inquiries thereupon* (Jo. Martyn and Ja. Allestry, 1665).

10.1126/science.adk1293

LETTERS

The Iberian wolf could soon lose some of the protection that has allowed the population to grow.

PHOTO: JUAN CARLOS MUNOZ/MINDEN PICTURES

Edited by **Jennifer Sills**

European Commission may gut wolf protection

The recovery of the wolf in Europe is one of the rare conservation successes on the continent (*1*). Instrumental to this recovery has been the strict legal protection of wolves throughout most of their range under Annex IV of the Habitats Directive (*2*). Strict protection has prevented anti-conservation interest groups from gaining the upper hand on wolf policy. The Court of Justice of the European Union has several times given a strict interpretation of the Habitats Directive in favor of wolf conservation (*3*). However, less than a year after it agreed to the ambitious Kunming-Montreal Global Biodiversity Framework, the European Commission has announced that it is considering a proposal to weaken the protection of wolves (*4*). This goes against a Directive "fitness check" study requested by the previous Commission, which concluded that changing annexes would be counterproductive and stressed the importance of enforcement (*5*).

The Commission has grown increasingly reluctant to fulfill its role of enforcing legal obligations for wolf conservation. It has failed to take action to prevent the repeated violations of EU law by Sweden for more than a decade (*6*) and has passively watched a wolf population disappear in Spain (*7*). It is now considering putting forward "a proposal to modify, where appropriate, the status of protection of the wolf," (*4*) presumably by moving the species from Annex IV to Annex V of the Habitats Directive, as demanded by farmer, landowner, and hunter organizations (*8*).

In some areas, wolves are already classified in Annex V, under which wolf killing does not need to be justified. In practice, Annex V listing means that there is very little oversight from the EU. The protection of Annex V species has sometimes been treated as optional. Finland, for instance, has for years considered Annex V listing of its wolves in the northern part of the country as a license to nearly eradicate them (*9*).

Although wolves under Annex V still need to have Favourable Conservation Status (FCS), as under Annex IV, the contentiousness of what constitutes FCS would leave ample room for Member States to set it at the lowest possible population size for political reasons. This is already the case in Sweden, where the government instructed its Environmental Protection Agency to set FCS to between 170 and 270 wolves (*10*).

A change in wolf protection would in practice mean far fewer restrictions on the killing of wolves in Europe, which is why Sweden and Austria have recently asked the Commission to implement it (*11*). Under Article 19 of the Habitats Directive, however, a change in wolf protection requires the unanimity of all 27 Member States (*12*). Conservationists in Europe therefore need to find a single unsupportive government to veto undoing decades of wolf recovery.

Guillaume Chapron[1]*, Yaffa Epstein[2,3], Mar Ouro Ortmark[2], Lovisa Helmius[1], Juan Pablo Ramírez Loza[1], Julien Bétaille[4], José Vicente López-Bao[5]
[1]Department of Ecology, Swedish University of Agricultural Sciences, 73993 Riddarhyttan, Sweden. [2]Department of Law, Uppsala University, 75120 Uppsala, Sweden. [3]Swedish Collegium for Advanced Studies, 75238 Uppsala, Sweden. [4]Law Faculty, Toulouse Capitole University, 31000 Toulouse, France. [5]Biodiversity Research Institute, Oviedo University, 33600 Mieres, Spain.
*Corresponding author.
Email: guillaume.chapron@slu.se

This work is the view of the authors and not necessarily that of the universities for which they work.

REFERENCES AND NOTES

1. G. Chapron *et al.*, *Science* **346**, 1517 (2014).
2. A. Trouwborst, F. M. Fleurke, *J. Int. Wildl. Law Pol.* **22**, 231 (2019).
3. Court of Justice of the European Union, Case C-674/17, ECLI:EU:C:2019:851 (2019).
4. European Commission, "Wolves in Europe: Commission urges local authorities to make full use of existing derogations and collects data for conservation status review" (2023); https://ec.europa.eu/commission/presscorner/detail/en/ip_23_4330.
5. Milieu Ltd., Institute for European Environmental Policy, ICF Consulting Services Ltd., "Evaluation Study to support the Fitness Check of the Birds and Habitats Directives" (2016).
6. L. Laikre *et al.*, *Science* **377**, 162 (2022).
7. R. Gamaza, "El lobo se ha extinguido en Andalucía," *El Correo de Andalucía* (2023); https://elcorreoweb.es/andalucia/el-lobo-se-ha-extinguido-en-andalucia-YD8702526 [in Spanish].
8. Committee of Professional Agricultural Organisations–General Confederation of Agricultural Cooperatives (Copa-Cogeca), European Landowners' Organization (ELO), European Council of Young Farmers (CEJA), Federation of Associations for Hunting and Conservation of the EU (FACE), "Joint Statement: The rise of large carnivore conflicts in Europe" (2022).
9. Finnish Ministry of Agriculture and Forestry, "Management plan for the wolf population in Finland" (2019) [in Finnish].
10. Swedish Government, "A Sustainable wolf policy: Government's proposition: 2012/13:191" (2013) [in Swedish].
11. Swedish Government, "Sweden and Austria agree to a joint action plan on EU wolf policy" (2023); https://www.regeringen.se/pressmeddelanden/2023/08/sverige-och-osterrike-enas-i-en-gemensam-handlingsplan-om-eus-vargpolitik/ [in Swedish].
12. European Union, "Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora" (2013).

COMPETING INTERESTS

G.C. is a member of the International Union for Conservation of Nature (IUCN) Species Survival Commission (SSC) Large Carnivore Initiative for Europe and a member of the French government–appointed Scientific Council for Wolf and Pastoralism, both unpaid advisory roles. M.O.O. was previously in the board of directors of the Association for the Conservation and Study of the Iberian Wolf, an unpaid advisory role. L.H. is a board member of the Sweden's Environmental Association of Law (Miljöjuristerna), an unpaid position. J.V.L.-B. is a member of the Canid Specialist Group of the IUCN SSC and a member of the Scientific Committee advising the Spanish Ministry for the Ecological Transition and the Demographic Challenge, both unpaid advisory roles.

10.1126/science.adk7686

Australia's undeveloped land is not "empty"

To address a housing crisis, Australia's federal government has set an ambitious target to build one million new houses within 5 years (*1*). However, clearing land

for construction will destroy habitats, escalating local species extinctions. Almost 2000 threatened species live in Australia, two-thirds of which are affected by land clearing (*2*). Instead of viewing open spaces as opportunities for development, Australia should embrace high-density urban planning to protect its natural habitats.

Australian extinctions represent 5 to 10% of global extinctions since the year 1500 (*3*), and the extinction trajectory is projected to worsen (*4*). One of the most biodiverse continents on Earth, Australia continues to extinguish endemic species (*5*) in part because its relatively small population fails to recognize the inherent value of its vast natural landscapes. For most of the past few hundred years, much of Australia was considered to be *terra nullius*, or "empty land," belonging to no one.

When Europeans first colonized Australia, the legal myth of *terra nullius* dispossessed Indigenous Australians of any claim to land ownership. By deeming the land "empty," colonialists justified taking it for themselves. In 1992, Indigenous Australian Edward Mabo sued the State of Queensland, arguing that Indigenous people had preexisting rights to the land (*6*). The High Court of Australia ruled in favor of Mabo, rejecting the doctrine of *terra nullius*. However, the idea that undeveloped land is "empty" continues to permeate Australia's land-use decisions. To protect biodiversity and cultural values, Australia must overcome the idea of land as void of life.

Every tract of land earmarked for development has cultural connection, is already occupied by countless species, and provides ecosystem services (*7*). The government's housing plans will further threaten many endemic species, including the iconic koala, as well as reduce Australia's ability to reach internationally agreed-upon targets on biodiversity goals (*8*) and climate commitments (*9*). Although it may seem counterintuitive, Australia needs to adopt urban planning solutions similar to those in countries with high population density. These solutions include transparent and systematic planning that avoids high biodiversity areas (*10*), efficient housing that reflects shrinking households, densification without compromising green spaces, and adaptive and innovative design to accommodate different uses (*11*, *12*). Providing safe and affordable housing for humans cannot occur at the cost of Australia's distinctive biodiversity and culture.

Elizabeth Brunton[1]*, Theresa Ashford[2], Romane H. Cristescu[1], Stefanie Fishel[2], Michelle Ward[3,4]
[1]School of Science Technology and Engineering, University of the Sunshine Coast, Sippy Downs, QLD, Australia. [2]School of Law and Society, University of the Sunshine Coast, Sippy Downs, QLD, Australia. [3]World Wide Fund for Nature–Australia, Brisbane, QLD, Australia. [4]School of Environment, University of Queensland, QLD, Australia.
*Corresponding author.
Email: ebrunton@usc.edu.au

REFERENCES AND NOTES

1. Commonwealth Government of Australia, "Improving housing supply and affordability" (2023); https://treasury.gov.au/housing-policy.
2. M. Ward *et al.*, *Ecol. Evol.* **11**, 11749 (2021).
3. J. C. Z. Woinarski *et al.*, *Biol. Conserv.* **239**, e108261 (2019).
4. H. M. Geyle *et al.*, *Pacif. Conserv. Biol.* **24**, 157 (2018).
5. S. Legge, L. Rumpff, S. G. Garnett, J. C. Z. Woinarski, *Science* **381**, 622 (2023).
6. High Court of Australia, "Mabo v Queensland (No 2) ('Mabo case') [1992] HCA 23; (1992) 175 CLR 1 (3 June 1992)" (1992); http://www.austlii.edu.au/cgi-bin/viewdoc/au/cases/cth/HCA/1992/23.html.
7. J. E. Engert, R. L. Pressey, V. M. Adams, *Conserv. Lett.* 10.1111/conl.12975 (2023).
8. K. Ma, *Biodivers. Sci.* **31**, 23133 (2023).
9. UN Framework Convention on Climate Change, "Paris Agreement (2015)," COP Report No. 21, Addendum (UN Doc. FCCC/CP/2015/10/Add, (2016).
10. S. Bekessey *et al.*, *Landsc. Urb. Plan.* **108**, 2 (2012).
11. S. E. Bibri, J. Krogstie, M. Kärrholm, *Dev. Built Environ.* **4**, 100021 (2020).
12. C. Haaland, C. Konijnendijk van Den Bosch, *Urb. For. Urb. Green.* **14**, 760 (2015).

10.1126/science.adk9055

Expanded US mentorship requires accountability

Early-career researchers are vital to the scientific enterprise but face challenges such as career uncertainty, tenuous support networks, and financial instability that decrease their interest in pursuing science, technology, engineering, and math (STEM) fields (*1*). High-quality mentorship can be essential to improving persistence in STEM, as well as academic performance, connectedness, and sense of belonging (*2*–*5*). The US National Science Foundation (NSF) has required grant applications to include a description of mentoring activities—called mentorship plans—to benefit postdoctoral researchers since 2007 (*6*). In 2022, the requirement for mentorship plans was expanded to include mentorship for US graduate students under the CHIPS and Science Act (*7*). However, neither the CHIPS and Science Act nor NSF requires a formal assessment of its mentorship plans by mentees, mentors, or independent assessors.

Every mentorship plan should include some measure of accountability to ensure that the next generation of scientists and engineers receive the support they need as they start their careers. Indeed, good models already exist within the agency. For example, NSF's Research and Mentoring for Postbaccalaureates (RaMP) program in Biological Sciences suggests that research proposals include an independent assessor, such as a member of a university center for teaching, to evaluate mentors and mentees. These assessments must occur throughout the mentorship (i.e., formative assessments) to allow for adjustments, as well as at the end of the mentorship (i.e., summative assessments) (*8*). The model allows those trained in assessing mentorship and familiar with specific mentor–mentee situations to conduct assessments, rather than leaving them to an NSF program officer's discretion.

Accountability for mentorship plans can be initiated by academic institutions, but it is crucial for federal funders to provide support. Now that graduate student mentorship plans are required for NSF-funded scientists, NSF can set the standard for best practices in mentorship (*9*). Ensuring accountability for mentorship plans through independent university assessors will be one option. Funders should also consider adding mentorship assessments to the data already required in annual and final reports, which are produced to show how grant funding was used.

Xoco Shinbrot[1]*, Julia Gerson[2], Brianne Gutmann[3], Kiki Ikossi[4], Lisa Ulmer[5]
[1]State of California, Delta Stewardship Council, Sacramento, CA, USA. [2]Avantiqor, Arlington, VAUSA. [3]Department of Physics and Astronomy, San Jose State University, San Jose, CA, USA. [4]George Mason University, Fairfax, VA, USA. [5]Patient-Centered Outcomes Research Institute, Washington, DC, USA.
*Corresponding author.
Email: x.shinbrot@gmail.com

The views and opinions expressed in this Letter are those of the authors and do not necessarily reflect the views of the United States Government. This Letter was written when the authors were fellows in the AAAS Science and Technology Policy Fellowships program. (AAAS is the publisher of *Science*.)

REFERENCES AND NOTES

1. R. Ysseldyk *et al.*, *Front. Psychol.* **10**, 1 (2019).
2. M. A. Burt, R. T. Barnes, S. Schanz, S. Clinton, E. V. Fischer, *Eos* **104**, 36 (2023).
3. S. W. Davies *et al.*, *PLOS Biol.* **19**, e3001282 (2021).
4. M. Estrada, P. R. Hernandez, P. W. Schultz, *CBE—Life Sci. Educ.* **17**, 1 (2018).
5. H. Haeger, C. Fresquez, *CBE—Life Sci. Educ.* **15**, 1 (2016).
6. Public Law 110-69, America Creating Opportunities to Meaningfully Promote Excellence in Technology, Education, and Science Act (2007); https://www.congress.gov/110/plaws/publ69/PLAW-110publ69.pdf.
7. Public Law 117-167, CHIPS Act of 2022 (2022); https://www.congress.gov/117/plaws/publ167/PLAW-117publ167.pdf.
8. National Science Foundation (NSF) 23-514, Research and Mentoring for Postbaccalaureates in Biological Sciences (RaMP) (2003); https://www.nsf.gov/pubs/2023/nsf23514/nsf23514.htm.
9. National Academies of Sciences, Engineering, and Math (NASEM), "Roundtable on Mentorship, Wellbeing, and Professional Development" (2023); https://www.nationalacademies.org/our-work/roundtable-on-mentorship-well-being-and-creativity-in-scientific-research.

10.1126/science.adk1696

RESEARCH

IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

EPISTASIS

When 2 plus 2 doesn't equal 4

Epistasis occurs when one gene or variant modifies the phenotypic effect of another, for example, when loss of an enzyme involved in pigmentation masks another gene's regulation of patterning. Aguirre *et al.* looked at a series of cis-regulatory alleles of three genes involved in meristem development, specifically examining the effect on tomato seed compartment number. They found that although some relationships between genes were dose dependent, many of these interactions had unexpected effects. These results not only add to our understanding of genetic interactions between three genes at once, but also give insight into the potentially unexpected phenotypes that can emerge from alleles engineered for agricultural purposes. —CNS *Science*, adi5222, this issue p. 315

Complex interactions between genes control fruit size and number of seed compartments in tomatoes.

SOLAR CELLS

Ultrathin and thermally stable SAMs

Inverted (p-i-n geometry) perovskite solar cells (PSCs) have fabrication advantages compared with conventional cells, but the self-assembled monolayers (SAMs) used to stabilize charge extraction layers have been prone to thermal degradation. Li *et al*. report on a phosphonic acid SAM that is anchored onto the particles in a nickel oxide film, optimizing the dipole moment for fast hole extraction and leading to a low defect density. For 1.53–electron volt p-i-n PSCs, a power conversion efficiency of 25.6% was achieved, and the cells maintained more than 90% of that efficiency after 1200 hours of operation at 65°C. —PDS

Science, ade9637, this issue p. 284

RADIO ASTRONOMY

A bright and distant fast radio burst

Fast radio bursts (FRBs) are brief flashes of radio emission from extragalactic sources. Ryder *et al.* detected a FRB and localized its source to a galaxy at a redshift of about 1, more than halfway back to the Big Bang. The burst is unusually bright, challenging models of the FRB emission mechanism. The authors also investigated how the intergalactic medium causes dispersion of the radio waves, finding more dispersion than expected from a correlation measured at lower redshift. They inferred the presence of magnetized plasma within the host galaxy, which has a complex morphology. —KTS

Science, adf2678, this issue p. 294

AVIAN ECOLOGY

Warming on farmland is worse for birds

Habitat transformation into agricultural and urban land uses is expanding and is accompanied by warming temperatures and increasing climate extremes. These dual stresses may interact to affect animal fitness because forests can buffer extreme temperatures more than open habitats such as agricultural lands. Using data from the citizen science program Project NestWatch, Lauck *et al*. examined how extreme heat influences birds' fledgling success in forests, grassland, developed areas, and agricultural lands across the continental US. They found different effects of extreme heat in different land uses. Birds nesting in agricultural lands had lower fledgling success at

extreme warm temperatures, but the opposite was true in forests. Future warming will likely negatively affect bird reproduction in human-dominated areas, especially among species of conservation concern. —BEL

Science, add2915, this issue p. 290

POLYMER CHEMISTRY

Building blocks for versatile plastics

Despite a broad range of appealing performance properties, polyolefins are difficult to break down and so contribute to an increasingly severe plastic waste problem. Zhao *et al*. report a class of polymers that behave similarly to polyolefins but are easier to deconstruct for recycling. Specifically, they prepared block copolymers from oligomers of straight and branched terminal diols, dehydrogenatively forming ester linkages between the hydrocarbon chains. Varying the block distribution tunes the properties of the resulting plastic, and hydrogenation can cleave the ester bonds to regenerate the building blocks. —JSY

Science, adh3353, this issue p. 310

ORGANIC CHEMISTRY

Biasing substitution at benzylic carbon

Nucleophilic substitution reactions are widely used to form bonds to carbon. In many cases, the reactions occur with precisely predictable stereochemistry, the spatial distribution of the substituents. However, certain carbon centers, such as those adjacent to aryl rings, are so reactive that the leaving group departs before the nucleophile arrives, the so-called S_N1 mechanism, resulting in random product stereochemistry. Singh *et al*. found that enveloping chiral anions could catalytically stabilize the benzylic cations that form in this S_N1 process at low temperature, thereby selectively controlling the stereochemistry of carbon–carbon, carbon–oxygen, and carbon–nitrogen bond formation. —JSY

Science, adj7007, this issue p. 325

ZOONOTIC DISEASES

Leishmaniaisis and the skin microbiome

Leishmaniasis is a zoonotic parasitic infection transmitted to humans through the bite of sandflies. It can manifest in several forms, including cutaneous leishmaniasis. Farias Amorim *et al*. examined the influence of the skin microbiome on disease outcomes in 62 patients infected with *Leishmania braziliensis*. Larger skin microbial burdens and overrepresentation of *Staphylococcus* spp. correlated with slower healing of cutaneous leishmaniasis lesions. Analysis of host and microbial transcripts revealed that lesions enriched for *S. aureus* expressed more inflammatory markers, especially interleukin-1β (IL-1β) family members. Cutaneous IL-1β delayed lesion healing in mice that had been colonized with *S. aureus* and infected with *L. braziliensis*. These findings reveal the influence that skin microbiota can have on the resolution of cutaneous leishmaniasis and highlight IL-1β as a potential immunotherapy target for cutaneous leishmaniasis. —CNF

Sci. Transl. Med. (2023) 10.1126/scitranslmed.adh1469

IMMUNOLOGY

Designing a cell-selective cytokine

The cytokine interleukin-2 (IL-2) activates cells that either perpetuate or suppress immune system activity, making cell selectivity critical for its use in immunotherapy. Orcutt-Jahns *et al*. used a model of binding and signaling dynamics to design a series of IL-2 variants with enhanced selectivity for immunosuppressive regulatory T cells. Versions that leveraged multivalent binding to the α chain of the heterotrimeric IL-2 receptor were more selective for regulatory T cells than were current IL-2 therapeutics. This approach may yield IL-2 variants that are more effective at treating autoimmune diseases. —LKF

Sci. Signal. (2023) 10.1126/scisignal.adg0699

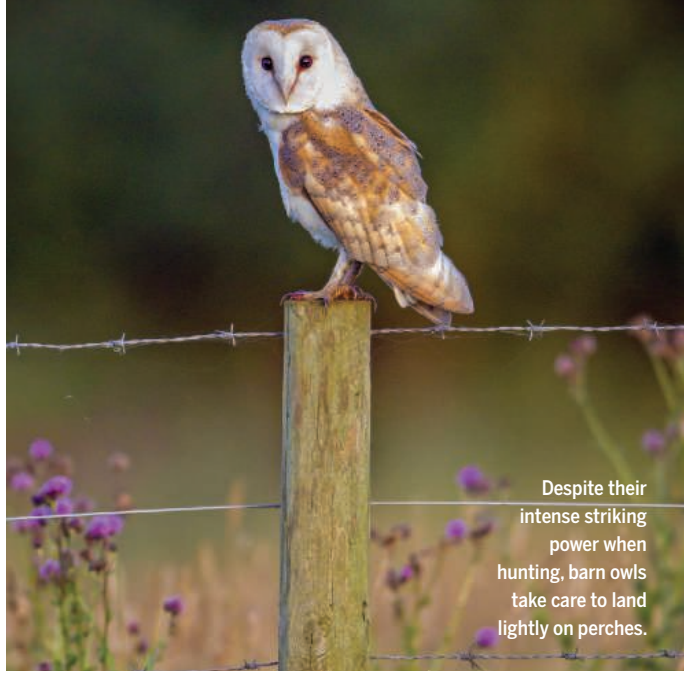
IN OTHER JOURNALS

Edited by **Caroline Ash** *and* **Jesse Smith**

ECOLOGY

Pale, silent, and deadly

Barn owls' otherworldly hunting flights excite awe in those people fortunate enough to see one, but even these ultrastealthy owls create turbulence and sound in flight, so many choose to perch silently and wait quietly for prey. Schalcher *et al.* monitored body movement and posture in relation to hunting results for 163 accelerometer- and GPS-tagged birds and found that the force of landing on perches predicted an owl's hunting success. Barn owls showed the highest strike force relative to body mass recorded for any bird so far (100 newtons). Fence boundary posts and other man-made objects provided the most successful perches because ease of access allowed better flight control and quiet landings. Human land use in this case actually benefits raptors by providing accessible and solid perches for stealthy hunting and maximum predation success. —CA *eLife* (2023) 10.7554/eLife.87775.1



Despite their intense striking power when hunting, barn owls take care to land lightly on perches.

NEUROSCIENCE

The mechanics of neurodevelopment

Spontaneous activity, characterized by transient increases in the concentration of calcium ions, are an important feature of neurons and glia in the developing nervous system. These calcium transients are thought to be a result of chemical signals such as neurotransmitters. However, in the developing zebrafish, Brandt and Smith found that calcium transients in satellite glia of the dorsal root ganglion were dependent on the activity of a mechanosensitive ion channel and could be elicited by mechanical stimulation of the tissue. —SAL

PLOS Biol. (2023) 10.1371/journal.pbio.3002319

PHOTO: PAUL SAWER/FLPA/SCIENCE SOURCE

ALSO IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

MEDICINE

New target for noncommunicable diseases

Bacteriophages selectively target bacteria and have a long history as antimicrobial therapy. The re-emergence of phage therapy for infectious diseases as an alternative to antibiotics has also increased interest in this approach to target bacteria that are associated with the emergence and progression of noncommunicable diseases (NCDs) such as inflammatory bowel disease, hepatitis, and some types of cancer. In a Perspective, Kviatcovsky *et al*. discuss the evidence for phage therapy in the treatment of NCDs; the possibilities of using phage-derived proteins and synthetic biology to broaden applications; and the challenges of overcoming resistance, personalizing phage selection, and the attendant regulatory implications. —GKA

Science, adh2718, this issue p. 266

CELL BIOLOGY

An in-depth look at LYTACs

Delivery to lysosomes of various biotherapeutics such as lysosomal enzyme replacement therapies, antibody-drug conjugates, and oligonucleotide therapeutics is crucial for their activity. Extracellular protein degraders, including lysosome-targeting chimeras (LYTACs), enable the precise depletion of proteins from the extracellular space through lysosomal targeting. Ahn *et al*. used unbiased genetic screens and proteomics to map cellular determinants of extracellular degrader activity. Activation of the E3 Cullin-3 enabled LYTAC target complexes to reach lysosomes and was a predictive marker for cell susceptibility to degraders. Cell surface CI-M6PR, a prototypical lysosome-targeting receptor, was found to be blocked by endogenous mannose-6-phosphate (M6P)–modified glycoproteins. The retromer complex counteracted this degradation activity through recycling LYTAC-CI-M6PR complexes. Accordingly, disruption of M6P biosynthesis or CI-M6PR recycling increased LYTAC activity. Detailed mechanistic study of membrane protein degradation by LYTACs can thus elucidate endocytic biology and could help in the rational deployment of degraders against disease. —SMH

Science, adf6249, this issue p. 281

STRATOSPHERIC OZONE

A big blast

The Hunga Tonga–Hunga Ha'apai eruption of 2022 injected huge amounts of water into the stratosphere and caused a large, rapid loss of ozone. Evan *et al*. collected in situ data on water, aerosols, and ozone in the volcanic plume and combined them with remote sensing observations to show that heterogeneous chlorine activation on humidified volcanic aerosols was the cause of the massive ozone loss that occurred. This loss was primarily triggered by the synergistic effects of strong humidification, radiative cooling, and added aerosol surface area, and this observation supports the suggestion that excess midlatitude stratospheric water associated with convection changes due to global warming could drive increases in lower stratospheric ozone loss. —HJS

Science, adg2551, this issue p. 282

NEUROSCIENCE

Different hippocampal coding strategies

Synchronous hippocampal neuronal ensemble activity supports episodic memory. This observation has led to the view that the main function of the hippocampus is to encode associations among different elements of an experience. However, an alternative hypothesis is that the hippocampus generates predictive representations of the world that can guide flexible behaviors. Liu *et al*. disrupted input from the entorhinal cortex to hippocampal area CA1 (see the Perspective by Steudler and Ólafsdóttir), thus destroying the sequence dynamics of place cells while keeping their coincidental firing intact. Sequence replay was disrupted but assembly reactivations were preserved. Different CA1 codes thus serve corresponding memory operations, with the place code supporting associative memory tasks and the sequence code supporting tasks that require learning about predictive transitions in space. —PRS

Science, adi8237, this issue p. 283; see also adk4642, p. 262

PHOTONICS

Unraveling hot-carrier dynamics

When electrons are optically excited high into the conduction band, they usually thermalize rapidly, losing energy through scattering processes. The ability to efficiently extract hot-carriers in the upper conduction band would be useful but is hampered by the lack of understanding of the processes involved. Taghinejad *et al.* developed a technique based on analyzing the terahertz pulses emitted by the hot-carriers (see the Perspective by Bangle and Mikkelsen). Modeling allowed them to unravel the spatiotemporal dynamics of the hot-carrier transport, providing a clearer picture for developing applications such as energy harvesting and photochemistry that can exploit these high-energy hot-carriers. —ISO

Science, adj5612, this issue p. 299; see also adk6862 p. 264

MARINE HEATWAVES

Mass starvation of snow crab

Marine heatwaves, a component of our impact on the Earth's climate, can bring both expected and unexpected environmental change. Between 2018 and 2021, after a period of historically high crab abundance and a series of marine heatwaves, the population of snow crab in the Bering Sea declined by 10 billion. Szuwalksi *et al*. used survey data to model the potential drivers of the decline in this ecologically and commercially important species. They found that the temperature of the water was not above the species' thermal limits, but it did increase their caloric needs considerably (see the Perspective by Kruse). This increase, in conjunction with a restriction in range, led to an unexpected mass starvation event. —SNV

Science, adf6035, this issue p. 306; see also adk7565 p. 260

FOOD WEB ECOLOGY

Cascading effects of cicadas

Periodical cicadas emerge en masse every 13 or 17 years, bringing not just a continuous drone of noise but also an influx of food for birds and other predators. Getman-Pickering *et al*. performed a set of experiments in the years before, after, and during the 2021 Brood X cicada emergence to determine how this addition of resources affected trophic dynamics in an eastern US forest (see the Perspective by Parker). They found that, compared with nonemergence years, cicada emergence led to lower caterpillar predation by birds, higher caterpillar densities, and lower herbivory rates on oak saplings. As birds opportunistically switched to cicada prey, their control over herbivory declined, showing the far-reaching

ALSO IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

short-term impacts of a resource pulse. —BEL

Science, adi7426, this issue p. 320; see also adk5880 p. 268

COMPUTER SCIENCE

A brain-inspired computer chip

The amount of data humans process and send around the globe on a daily basis is astonishing. However, the energy cost involved is high, and there is a strong need for designing energy-efficient devices. Modha *et al*. describe a chip with a neural inspired architecture, called NorthPole, that achieves substantially higher performance, energy efficiency, and area efficiency compared with other comparable architectures (see the Perspective by Iyer and Roychowdhury). A key feature of this chip is the recognition that for almost all kinds of computing, access to memory plays as important a role as logic processing. Unlike analog in-memory computing, this purely digital system has the option of tailoring the bit precision as needed, which allows for optimization of the power usage. —PRS

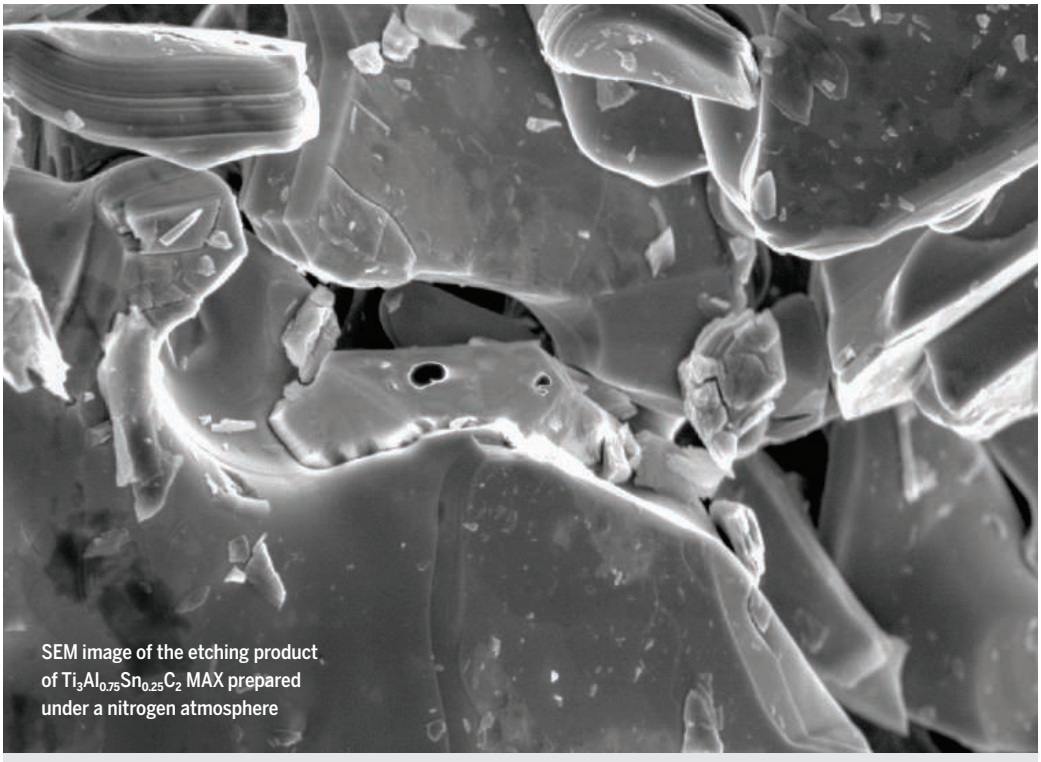
Science, adh1174, this issue p. 329; see also adk6874 p. 263

NEUROIMMUNOLOGY

Restraining inflammation

The cytokine interleukin-31 (IL-31) has emerged as a key player in multiple allergic diseases. For example, in atopic dermatitis, IL-31 is believed to be a strong driver of itch, but the interplay between its activity on sensory neurons and in type 2 inflammation remains unclear. In a mouse model of chronic house dust mite–induced dermatitis, Fassett *et al.* found that IL-31 restrained cutaneous type 2 inflammation by initiating a pathway of neurogenic inflammation. Mice lacking the IL-13 receptor gene in sensory neurons displayed enhanced accumulation of type 2 cytokine–producing $CD4^+$ T cells, which were suppressed by the IL-31–induced neuropeptide calcitonin gene-related protein. Together, these findings demonstrate that in addition to triggering pruritus, IL-31 acts on sensory neurons to restrain type 2 inflammation in the skin during chronic allergen exposure. —CO

Sci. Immunol. (2023) 10.1126/sciimmunol.abi6887



SEM image of the etching product of $Ti_3Al_{0.75}Sn_{0.25}C_2$ MAX prepared under a nitrogen atmosphere

NANOMATERIALS

Leaving some A on MXenes

Layered MXene materials, in which M is an early transition metal and X is carbon or nitrogen, are made from MAX phases, where an A group element (usually aluminum) is removed by etching. Bagheri *et al.* used a mixed tin–aluminum composition for the A site and made an MXene in which tin atoms uniformly decorated the MXene surfaces. The presence of tin made the $Ti_3Al_{0.75}Sn_{0.25}C_2$ precursor less reactive, and a lithium intercalation step was needed before hydrofluoric acid etching could remove the aluminum. These MXene sheets had enhanced environmental stability, which was attributed to tin atoms passivating defect sites where oxidation could occur. —PDS *ACS Nano* (2023) 10.1021/acsnano.3c02198

CELL BIOLOGY

To be or not to be a stress kinase inhibitor?

The integrated stress response (ISR) is a vital cellular signaling pathway and is the focus of many drug development programs. The kinases that activate the ISR, including PERK and GCN2, are prominent drug targets, and various inhibitors have been generated for them. These inhibitors are widely used as tool compounds to evaluate the function of the ISR and the potential therapeutic benefit of inhibiting it under a range of conditions. Unexpectedly, Szaruga *et al*. found that ATP-competitive inhibitors of the ISR-activating kinases can paradoxically activate the stress response pathway by directly binding to and activating a sister ISR kinase. —SMH

Nat. Commun. (2023) 10.1038/s41467-023-40823-8

CASH TRANSFERS

Money well spent

Two percent of the world remains unhoused, and lack of money is a primary cause of homelessness. However, governments traditionally support unhoused people through emergency health care and temporary housing, services that do not empower people to make their own financial decisions. Although unconditional cash transfers have been effective in lower-income countries, we know little about their impact on unhoused individuals in affluent countries. To bridge this knowledge gap, Dwyer *et al.* studied unhoused individuals in Vancouver, Canada, and offered them a one-time Canadian $7500 cash transfer. Compared with control participants, over the course of a year, cash recipients spent fewer days unhoused, more days in stable housing, and retained more savings without reckless spending on nonessential goods. This finding challenges public perceptions and policies assuming that unhoused individuals lack the money management skills necessary to become financially independent. —EEU

Proc. Natl. Acad. Sci. U.S.A. (2023) 10.1073/pnas.2222103120

MACHINE LEARNING

Toward next-generation lithium batteries

Recent advances in computer science and computational chemistry may enable useful insights into poorly understood complex chemical systems, for which missing databases can be constructed by combining first-principles methods with machine learning models. Using graph theory–based algorithms and density functional theory calculations, Gao *et al.* generated an extensive database of potential electrolyte solvent molecules for lithium metal batteries. Analysis of this database with Shapley additive explanations based on the random forest model revealed the key factors influencing the reductive stability of electrolyte solvents, which could be used to accelerate the rational design of electrolyte molecules for these promising but largely unexplored energy storage systems. —YS

J. Am. Chem. Soc. (2023) 10.1021/jacs.3c08346

RESEARCH POLICY

Avoiding a race to the bottom

When grant funders allocate resources across heterogeneous fields based on the number of applications in each field, "noisier" fields in which views of what constitutes good science are more divergent can attract more and weaker applications, driving applications and funding away from fields for which quality is more consistently assessed. Based on their study of 2014 reforms in the European Research Council funding process, Adda and Ottaviani suggest that funders consider reforms such as pooling applications across fields. These findings may also apply to journals evaluating manuscripts and universities admitting students. —BW

Q. J. Econ. (2023) 10.1093/qje/qjad046

RESEARCH ARTICLE SUMMARY

CELL BIOLOGY

Elucidating the cellular determinants of targeted membrane protein degradation by lysosome-targeting chimeras

Green Ahn, Nicholas M. Riley, Roarke A. Kamber, Simon Wisnovsky, Salvador Moncayo von Hase, Michael C. Bassik, Steven M. Banik*, Carolyn R. Bertozzi*

INTRODUCTION: Repurposing endogenous cellular machinery to degrade disease-driving proteins is a promising therapeutic strategy, as evidenced by the rapid development of intracellular degraders over the past two decades. Extracellular degraders, including lysosome-targeting chimeras (LYTACs), have expanded the scope of protein degradation to secreted and plasma membrane targets. LYTACs harness endogenous cell surface receptors, such as cation-independent mannose 6–phosphate receptor (CI-M6PR), to traffic specific proteins between the plasma membrane and lysosomes. However, the cellular features that affect the ability to hijack lysosomal trafficking for membrane protein degradation are largely unknown. An approach to map features that potentiate or attenuate the ability to degrade membrane proteins is needed to further understand extracellular degradation mechanisms.

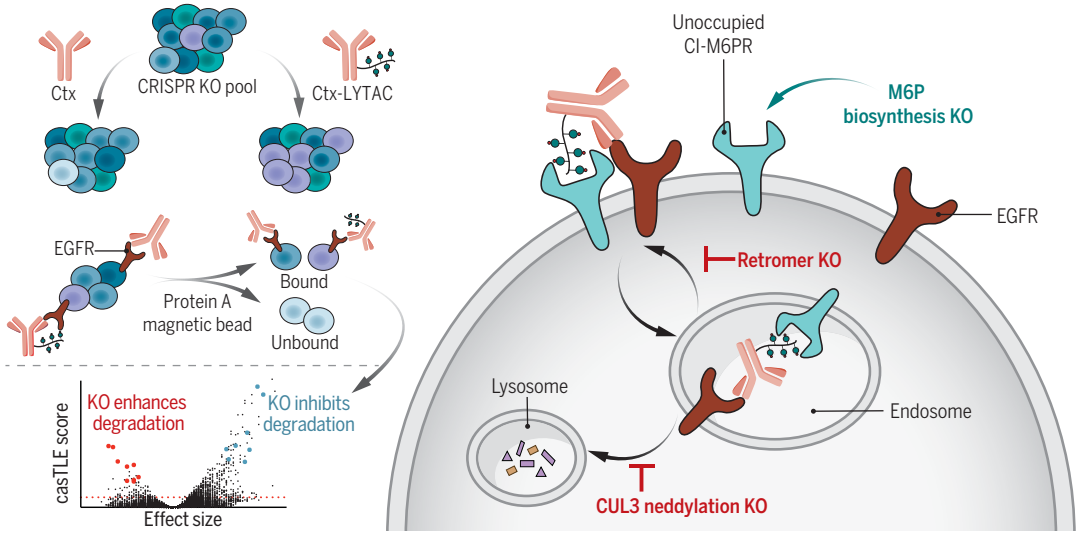
RATIONALE: The factors that affect membrane protein degradation could facilitate rational therapeutic development and reveal biological insights crucial for numerous therapeutic modalities. Membrane proteins, unlike soluble secreted proteins, can retain their own distinct trafficking behavior through the endosomal system, which must be overcome or enhanced for targeted degradation by extracellular degraders. We deploy an unbiased genome-wide CRISPR knockout screening approach complemented by proteomics to identify and characterize cellular components that modulate LYTAC-mediated membrane protein degradation. This approach combines modular chemically defined LYTAC ligands, scalable cell separation through magnetic beads, immunoprecipitation, and cell surface proteomics to identify key molecular players.

RESULTS: This genome-wide screen confirmed the importance of several endolysosomal genes for LYTAC-mediated degradation. We characterized the effect of additional genes distinct from the canonical endolysosomal pathway or genes whose knockout can enhance LYTAC activity. We identified that the retromer complex recycles LYTAC–CI-M6PR complexes from endosomes to the cell surface before dissociation of LYTAC from the receptor. Disruption of retromer complex genes enhanced degradation mediated by LYTAC by inhibiting this competitive recycling pathway. Genes involved in the neddylation of cullin 3 (CUL3), and therefore activation of its E3 ligase activity, were crucial for the trafficking of LYTAC-target complexes to lysosomes. We determined that levels of neddylated C[illegible] are correlated with degradation efficacy, th[illegible] by serving as a potential predictive marker for LYTAC activity. In addition, we revealed that genes involved in the mannose 6–phosphate (M6P) biosynthesis pathway counteract LYTAC activity through synthesizing competitive ligands for CI-M6PR. We observed that cell surface CI-M6PR receptors are occupied by distinct M6P-modified lysosomal glycoproteins. Disruption of M6P biosynthesis increased the fraction of unoccupied receptors on the cell surface, allowing enhanced LYTAC-receptor engagement and internalization.

CONCLUSION: Mechanistic studies driven by extracellular degraders can reveal previously unknown aspects of receptor and endosomal biology. This study uncovered drivers of extracellular degradation using CI-M6PR as a lysosome trafficking receptor. The findings of this study motivate new design strategies for lysosome-targeted protein degraders. Furthermore, the magnetic enrichment–based genomic screen approach can be expanded to any other extracellular degraders, allowing the identification of common and distinct mechanisms across extracellular degradation platforms. We anticipate that the impacts of receptor recycling, occupancy, and ability to traverse the endolysosomal system revealed in our work will have broader implications for other therapeutics that rely on cell surface–to–lysosome trafficking, such as enzyme-replacement therapies, nucleic acid delivery, and antibody-drug conjugates. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: bertozzi@stanford.edu (C.R.B.); sbanik@stanford.edu (S.M.B.)
Cite this article as G. Ahn *et al.*, *Science* **382**, eadf6249 (2023). DOI: 10.1126/science.adf6249

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adf6249



Identifying cellular determinants of LYTAC-mediated degradation of membrane proteins. A genome-wide CRISPR screen was conducted to identify factors regulating LYTAC-mediated degradation of epidermal growth factor receptor (EGFR) via magnetic cell sorting. Knockout (KO) of retromer complex resulted in reduced recycling of LYTAC-target complex. CUL3 neddylation was essential for lysosomal trafficking and degradation. Disruption of M6P biosynthesis pathway resulted in increased unoccupied CI-M6PR on the cell surface, enhancing LYTAC-receptor internalization. Ctx, cetuximab.

RESEARCH ARTICLE

CELL BIOLOGY

Elucidating the cellular determinants of targeted membrane protein degradation by lysosome-targeting chimeras

Green Ahn[1,2], Nicholas M. Riley[1,2], Roarke A. Kamber[3], Simon Wisnovsky[4], Salvador Moncayo von Hase[1,2], Michael C. Bassik[2,3], Steven M. Banik[1,2]*, Carolyn R. Bertozzi[1,2,5]*

Targeted protein degradation can provide advantages over inhibition approaches in the development of therapeutic strategies. Lysosome-targeting chimeras (LYTACs) harness receptors, such as the cation-independent mannose 6–phosphate receptor (CI-M6PR), to direct extracellular proteins to lysosomes. In this work, we used a genome-wide CRISPR knockout approach to identify modulators of LYTAC-mediated membrane protein degradation in human cells. We found that disrupting retromer genes improved target degradation by reducing LYTAC recycling to the plasma membrane. Neddylated cullin-3 facilitated LYTAC-complex lysosomal maturation and was a predictive marker for LYTAC efficacy. A substantial fraction of cell surface CI-M6PR remains occupied by endogenous M6P-modified glycoproteins. Thus, inhibition of M6P biosynthesis increased the internalization of LYTAC-target complexes. Our findings inform design strategies for next-generation LYTACs and elucidate aspects of cell surface receptor occupancy and trafficking.

Selective depletion of individual proteins is a validated approach for modulating numerous disease-driving targets. Intracellular degradation approaches using proteolysis-targeting chimeras (PROTACs) (*1*, *2*) or molecular glues (*3*, *4*), both of which recruit E3 ligases to promote ubiquitination of target proteins, have already demonstrated clinical impact (*5*, *6*). Technologies for targeted degradation of extracellular proteins, including lysosome-targeting chimeras (LYTACs) (*7*, *8*), molecular degraders of extracellular proteins through the asialoglycoprotein receptor (MoDEAs) (*9*), antibody-based PROTACs (AbTACs) (*10*), proteolysis-targeting antibodies (PROTABs) (*11*), and cytokine receptor–targeting chimeras (KineTACs) (*12*), have recently expanded the scope of potential therapeutic targets. A common theme of these approaches is harnessing the cellular internalization machinery to drive extracellular protein degradation in the lysosome. For example, LYTACs are bifunctional molecules that form ternary complexes with a protein of interest and a lysosome trafficking receptor, such as the cation-independent mannose 6–phosphate receptor (CI-M6PR) (*7*) (Fig. 1A) or the liver-restricted asialoglycoprotein receptor (ASGPR) (*8*, *13*).

[1]Department of Chemistry, Stanford University, Stanford, CA 94305, USA. [2]Sarafan ChEM-H, Stanford University, Stanford, CA 94305, USA. [3]Department of Genetics, Stanford University, Stanford, CA 94305, USA. [4]Pharmaceutical Sciences, University of British Columbia, Vancouver, BC, Canada. [5]Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305, USA.
*Corresponding author. Email: bertozzi@stanford.edu (C.R.B.); sbanik@stanford.edu (S.M.B.)

The development of PROTACs and molecular glues has been accelerated by mechanistic studies that have revealed key cellular determinants for induced protein degradation (*14*–*16*). However, the cellular features that enable or inhibit membrane protein degradation remain unclear. An understanding of the mechanisms driving targeted degradation of extracellular and transmembrane proteins could potentially accelerate therapeutic development and application. Given the importance of lysosomal trafficking for the delivery of various biotherapeutics, including lysosomal enzyme replacement therapies (ERTs) (*17*, *18*), antibody-drug conjugates (ADCs) (*19*, *20*), and oligonucleotide therapeutics (*21*–*23*), characterizing the mechanisms of extracellular degraders is likely to reveal principles that can benefit numerous therapeutic modalities. In this work, we performed a genome-wide CRISPR knockout (KO) screen to identify essential determinants of CI-M6PR–LYTAC–mediated lysosomal trafficking and degradation of membrane proteins.

Design of next-generation LYTACs comprising antibody-glycopeptide conjugates

To initiate a mechanistic study, we first developed synthetically scalable next-generation LYTAC molecules. Synthesis of first-generation CI-M6PR LYTACs used NCA (*N*-carboxyanhydride) polymerization to generate mixtures of alanine–serine–mannose 6–phosphonate (M6Pn) copolymers with various lengths and compositions (*7*). The heterogeneity of the initial M6Pn-glycopolymer could result in batch-to-batch differences and prevented characterization of the precise molecular nature of the LYTAC molecule. We developed homogeneous peptides to maintain control over structural composition of the antibody-peptide conjugates. We established a synthetic route to synthesize Fmoc-Serine(M6Pn)-OH monomer that could be easily incorporated into a standard solid-phase peptide synthesis (SPPS) workflow with high yield (fig. S1). Glycopeptides bearing two M6Pn ($M6Pn_2$) and five M6Pn ($M6Pn_5$) residues were readily conjugated to cetuximab (Ctx), the clinically approved anti–epidermal growth factor receptor (EGFR) antibody, using Copper-free click chemistry (Fig. 1B and fig. S2). Ctx-$M6Pn_5$ and Ctx-$M6Pn_2$ degraded ~80% of cell surface EGFR in UMRC2 and HeLa cells (Fig. 1C and fig. S3A). Maximum degradation was achieved at 10 nM for both Ctx-$M6Pn_5$ and Ctx-$M6Pn_2$, and higher concentrations did not reduce degradation (fig. S3C). More than 50% of the total EGFR was degraded after 24 hours of treatment in UMRC2 and HeLa cells (fig. S3, D and E). The presence or absence of serum did not affect the degradation efficacy (fig. S3F). We used a fluorophore-labeled rabbit immunoglobulin G (IgG) as a cargo in the presence of goat-anti-rabbit (control) or goat-anti-rabbit-$M6Pn_5$ (LYTAC) to evaluate internalization of a soluble protein. The half maximal inhibitory concentration (IC_{50}) for monomeric M6P inhibition of IgG uptake by goat-anti-rabbit-$M6Pn_5$ was 4.3 mM for LYTAC-mediated internalization (fig. S3G), and saturation of cargo uptake was reached at 100 nM LYTAC (fig. S3H).

To expand the targetable membrane protein scope, we conjugated the peptides to girentuximab (Gir) and onartuzumab (Ona)—antibodies against carbonic anhydrase IX (CA9) and c-Met, respectively. Gir-$M6Pn_5$ and Gir-$M6Pn_2$ degraded >80% of surface CA9 in UMRC2 cells and 70 to 80% in U87MG cells (Fig. 1D and fig. S3B). Only minor differences in efficacy were observed between $M6Pn_2$ and $M6Pn_5$, although Gir-$M6Pn_2$ conjugates were optimal for CA9 degradation, and Ctx-$M6Pn_5$ conjugates were optimal for EGFR degradation. We confirmed that Gir-$M6Pn_2$ degraded CA9 (Fig. 1E) and demonstrated that the down-regulation was highly selective for CA9 by quantitative proteomics (Fig. 1F). Given that only a small percentage of CI-M6PR is recycled to the plasma membrane, we investigated whether there is a capacity limit to CI-M6PR–induced degradation by assessing whether we could degrade two different membrane targets simultaneously. Ona-$M6Pn_5$ drove degradation of c-MET, and cotreatment with Ona-$M6Pn_5$ and Ctx-$M6Pn_5$ resulted in coincident degradation of EGFR and c-MET in two cell lines (Fig. 1G and fig. S4). We did not observe a loss of degradation efficacy with dual treatments, which suggests that the capacity for CI-M6PR–mediated internalization is not saturated through LYTAC appropriation.

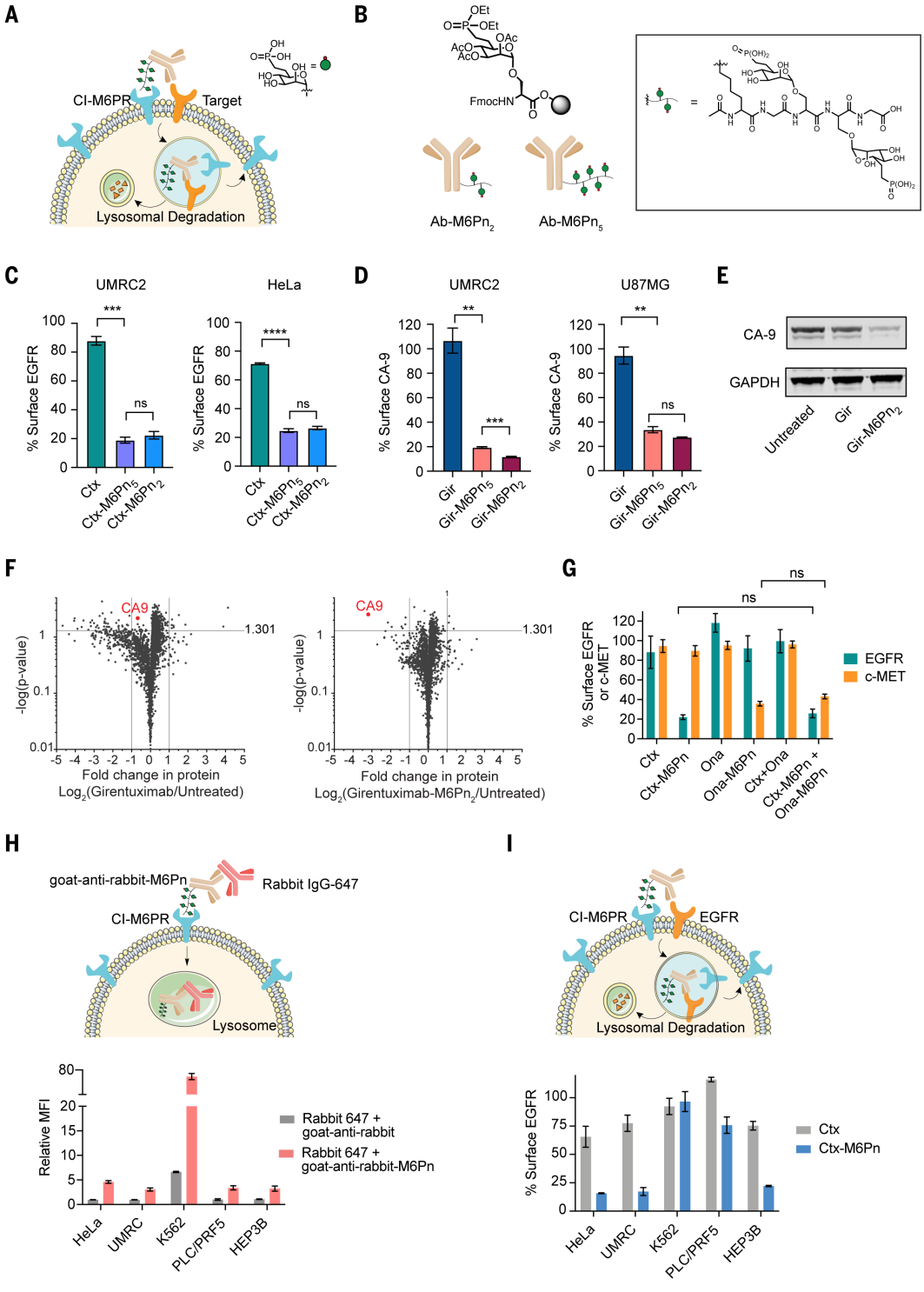


Fig. 1. LYTACs comprising antibody-glycopeptide conjugates enable degradation of membrane targets. (**A**) Membrane protein degradation mediated by LYTACs that harness CI-M6PR. (**B**) Synthesis of homogeneous M6Pn ligands through solid-phase peptide synthesis (SPPS). (**C**) Degradation of EGFR with M6Pn-LYTACs in UMRC2 and HeLa cells as determined by live-cell flow cytometry after 48 hours of treatment with 10 nM Ctx or LYTACs. (**D**) Degradation of CA9 in UMRC2 and U87MG cells as determined by live-cell flow cytometry after 48 hours of treatment with 10 nM Gir or LYTACs. (**E**) Immunoblot analysis of CA9 levels in UMRC2 cells after treatment with 10 nM Gir or Gir- $M6Pn_2$ for 48 hours. GAPDH, glyceraldehyde-3-phosphate dehydrogenase. (**F**) Fold change in the abundance of proteins in UMRC2 cells detected by quantitative proteomics analysis after 48 hours of treatment with 10 nM Gir or Gir-$M6Pn_2$. (**G**) Degradation of cell surface c-MET and EGFR in UMRC2 cells as determined by live-cell flow cytometry after 48 hours of single treatment with 10 nM Ctx or Ona-M6Pn or cotreatment of Ctx and Ona-M6Pn. (**H**) LYTAC-mediated uptake of rabbit IgG-647 in various cell lines. Mean fluorescence intensity (MFI) relative to the control (rabbit IgG-647 only) for cells incubated at 37°C for 1 hour with 50 nM rabbit IgG-647 and 25 nM goat anti-rabbit or goat anti-rabbit M6Pn. MFI was determined by live-cell flow cytometry. (**I**) Degradation of cell surface EGFR in various cell lines as determined by live-cell flow cytometry after 48 hours of treatment with 10 nM Ctx or Ctx conjugates. For (C), (D), and (F) to (I), data represent three independent experiments, and data are shown as means ± SEMs. *P* values were determined by unpaired two-tailed *t* tests. NS, not significant; **P* ≤ 0.05; ***P* ≤ 0.01; ****P* ≤ 0.001; *****P* ≤ 0.0001.

We evaluated soluble cargo internalization using fluorophore-labeled rabbit IgG with a $M6Pn_5$-conjugated goat-anti-rabbit (LYTAC) across five cell lines. K562 had significantly higher LYTAC-mediated internalization, whereas the other cell lines demonstrated low to moderate uptake capacity (Fig. 1H). The trend for internalization of soluble cargo correlated with cell surface CI-M6PR expression levels across the cell lines (fig. S5). We asked whether successful uptake of soluble cargo predicted successful degradation of membrane proteins. Although HeLa, UMRC2, and HEP3B showed efficient surface degradation of EGFR, they exhibited poor soluble cargo internalization capacity. No EGFR degradation was observed upon LYTAC treatment in K562 cells (Fig. 1I). The lack of correlation between the uptake of soluble cargo and membrane protein degradation suggests that regulators of induced membrane protein degradation are different

from those for soluble cargo uptake. A previous CRISPRi screen had determined essential genes for uptake of soluble cargo (NeutrAvidin-647) by biotinylated poly(M6Pn) (*7*). Given the disconnect between soluble protein uptake and membrane protein degradation efficiency, we reasoned that features that drive the two processes might differ. Membrane proteins play critical roles in tumorigenesis and neurological diseases (*24*, *25*), and the ability to degrade them in various contexts would be a beneficial therapeutic approach. A mechanistic understanding of induced membrane protein degradation could expand and generalize the technology while providing insight into endocytic regulators.

A magnetic cell sorting–based genome-wide CRISPR screen identifies key regulators of LYTAC-mediated degradation

Magnetic cell sorting (MACS) approaches enable rapid sorting of adherent cells while maintaining high cell viability (*26*–*28*). To provide insight into membrane protein degradation, we designed a genome-wide CRISPR deletion screen using magnetic enrichment. We generated UMRC2-Cas9 cells infected with a genome-wide single-guide RNA (sgRNA) knockout library (*29*) and treated cells with Ctx or Ctx-M6Pn. Populations that lost EGFR expression on the cell surface were isolated by sequestering EGFR-positive cells with protein A magnetic beads, which bind to the Fc region of human antibodies (Ctx). No differences in protein A binding between Ctx and Ctx-M6Pn were observed. Cells treated with Ctx-M6Pn that lost the ability to degrade EGFR through CRISPR-mediated gene knockout were retained in the bound fraction, whereas cells that lost cell surface EGFR remained in the unbound fraction. These enriched fractions were analyzed by next-generation sequencing to identify the genes required for LYTAC-mediated membrane protein degradation (Fig. 2A). Analysis and comparison of the guide compositions were performed using casTLE (*30*). Genes with positive effect size potentiate degradation and have higher sgRNA enrichment in the bound fraction. Genes with negative effect size attenuate degradation and have less abundant sgRNAs in the bound fraction (Fig. 2B). *IGF2R*, an alternative gene name for CI-M6PR, was a top hit with a positive effect size from the Ctx-M6Pn screen, as expected because M6Pn-based LYTACs rely on the binding and trafficking of CI-M6PR. Ctx alone marginally down-regulated EGFR, but the genes that drive Ctx-mediated EGFR internalization were distinct from those found for EGFR degradation mediated by Ctx-M6Pn (fig. S6). Gene ontology (GO) analysis for the hits from either Ctx or LYTAC treatment differentiated the cellular pathways used by these molecules. Enriched sgRNAs from the LYTAC treatment showed cellular components and biological processes involved in endocytosis, whereas sgRNAs from the Ctx treatment used cellular components related to known Ctx-mediated trafficking and nuclear localization of EGFR. We found "retromer complex," "CUL3-RING ubiquitin ligase," and "protein neddylation" as enriched specific pathways for LYTAC-mediated EGFR degradation (fig. S7, A and B). Mapping the positive effect hits from LYTAC treatment into subcellular components identified several endolysosomal pathway drivers such as "clathrin-coated pit," "early endosome," "late endosome," and "lysosome," mirroring the trafficking of CI-M6PR. Because these hits are known regulators of endocytosis, we sought to investigate the genes outside of this category, such as genes distinct from the endolysosomal pathway or genes whose knockout can enhance LYTAC-mediated membrane protein degradation (Fig. 2C). Genes associated with the retromer complex, neddylation pathway, and M6P biosynthesis in the endoplasmic reticulum (ER) and Golgi all had significant effects on LYTAC-mediated degradation but without clear mechanistic bases.

Disruption of retromer complex genes enhances LYTAC-mediated degradation by reducing recycling from the endosomes to the plasma membrane

The retromer complex is involved in retrograde transport from endosomes to the trans Golgi network (TGN) or the plasma membrane. Core components of the retromer complex, such as vacuolar protein sorting–associated protein 35 (VPS35), vacuolar protein sorting–associated protein 26A (VPS26A), sorting nexin 3 (SNX3), and vacuolar protein sorting–associated protein 29 (VPS29), were identified as negative hits (Fig. 3A, fig. S8, and table S1) from our screen. The retromer complex is responsible for retrograde transport of CI-M6PR through engagement with sorting signals within the cytosolic tail of CI-M6PR (*31*–*33*). Given that retromer genes had negative effect sizes, we questioned whether knockout of these genes would improve LYTAC-mediated degradation. We observed significantly greater degradation of EGFR in VPS26A- or SNX3-deficient cells after 3-hour and 48-hour treatment with LYTAC, which demonstrates that degradation is more efficient when retromer genes are disrupted (Fig. 3, B and C). We achieved >90% degradation of cell surface EGFR in VPS26A-deficient cells. In addition, confocal imaging revealed that EGFR was localized in both late endosomes or lysosomes (LE/LY) and at the plasma membrane of wild-type (WT) cells, whereas EGFR was mostly undetectable on the cell surface of VPS26A-deficient cells after LYTAC treatment (fig. S9A). Loss of VPS26A did not change the surface expression level of CI-M6PR (Fig. 3D). In LYTAC-treated cells, total CI-M6PR levels remained unchanged in both WT and VPS26A-deficient cells, whereas EGFR levels were reduced in both cell types, which confirms that LYTAC degrades its target without degrading CI-M6PR (fig. S9B). Furthermore, no change in LYTAC-mediated internalization of rabbit IgG was observed in VPS26A-deficient cells at 10, 20, and 60 min (Fig. 3E and fig. S9C).

Because the retromer complex is involved in recycling cargo from endosomes, it may partially recycle the LYTAC–CI-M6PR complex before dissociation of LYTAC from the receptor. To test this, we performed a pulse-chase experiment where we treated with Ctx-M6Pn for 24 hours and then washed and incubated with fresh media for an additional 24 hours. The localization of Ctx-M6Pn (LYTAC) was then visualized by goat-anti-human-647 staining. In WT cells, LYTAC was localized at both the plasma membrane and partially in LE/LY, whereas we did not observe LYTAC localization on the cell surface in VPS26A KO cells (Fig. 3F and fig. S10). Using flow cytometry, we observed an increase in cell surface staining of LYTAC with anti-human-647 in WT cells over time, whereas no increase in staining was observed in VPS26A-deficient cells (Fig. 3G). Thus, LYTACs can recycle back to the plasma membrane from endosomes via the retromer complex, whereas such recycling is reduced when a component in the retromer complex is disrupted. To identify whether the recycling of LYTAC was because of recycling of EGFR or CI-M6PR, we conducted the same pulse experiment with Ctx and observed that Ctx was still bound to EGFR on the cell surface in both WT and VPS26A-deficient cells (Fig. 3, F and G). Thus, the retromer complex can recycle the LYTAC–CI-M6PR complex before dissociation of LYTAC from the receptor, and perturbation of retromer genes can increase the efficacy of LYTAC by removing a competitive recycling pathway of CI-M6PR.

Neddylation of CUL3 is essential and a predictive marker for LYTAC-mediated membrane protein degradation

We found several genes with high casTLE scores that are involved in neddylation and activation of the E3 ligase cullin-3 (CUL3). During the activation process, Nedd8 is transferred from E1 [Nedd8 activating enzyme E1 subunit 1 (NAE1), ubiquitin-activating enzyme 3 (UBA3)] to E2 [Nedd8-conjugating enzyme Ubc12 (UBE2M)], then to E3 ligase (CUL3). Inactive CUL3 is normally bound by cullin-associated Nedd8-dissociated protein 1 (CAND1), and upon neddylation, CAND1 dissociates activating CUL3. CUL3 neddylation can be reversed by the constitutive photomorphogenesis 9 (COP9) signalosome complex to allow recycling of neddylation (*34*–*36*) (Fig. 4A and fig. S11A). All of these components, including CUL3, were

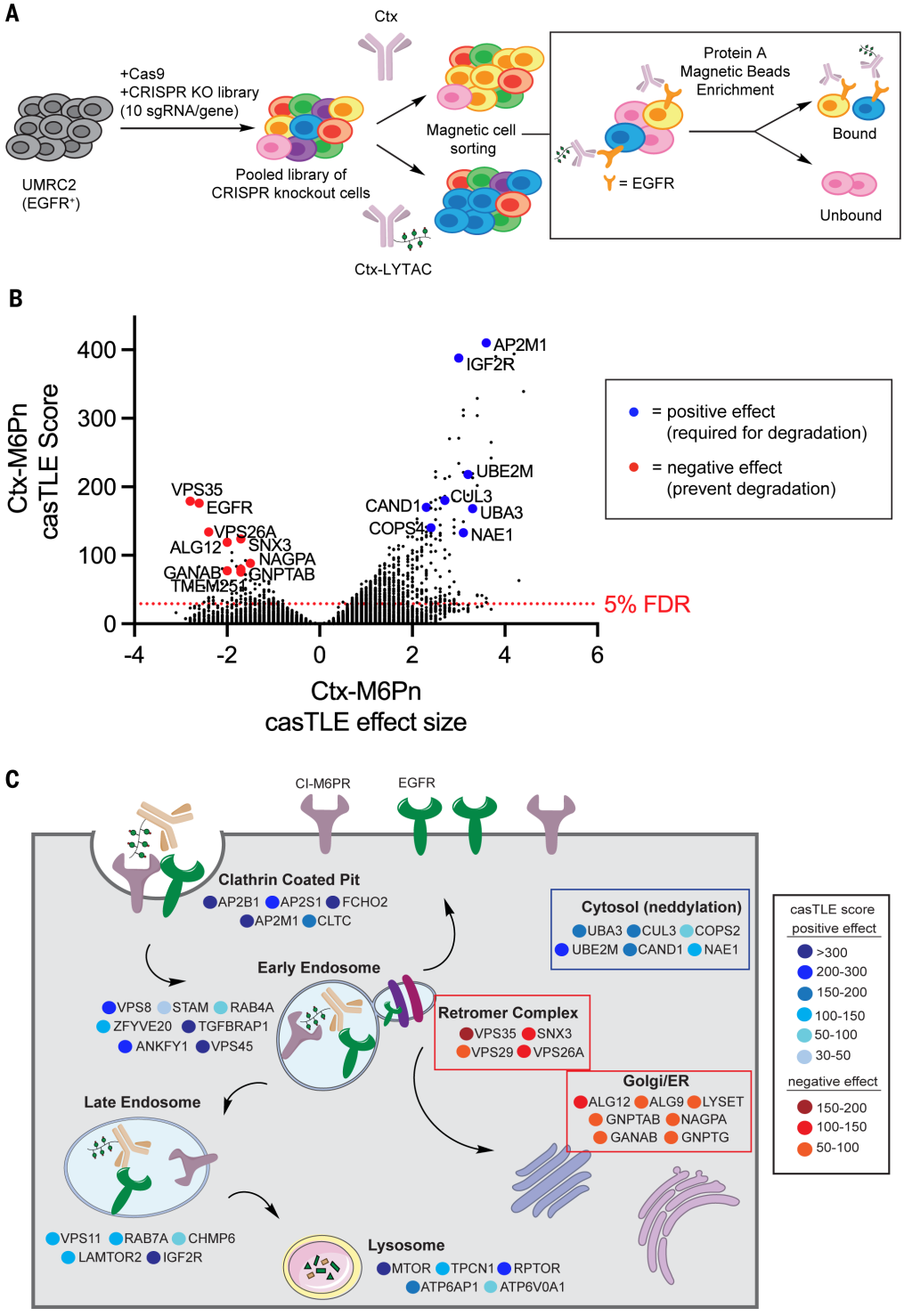


Fig. 2. Genome-wide CRISPR KO screen identifies regulators for LYTAC-mediated membrane protein degradation. (**A**) MACS-based CRISPR KO screen in UMRC2 cells stably expressing Cas9 and a library of sgRNAs with a genome-wide coverage. Cells were treated with Ctx or Ctx-LYTAC for 48 hours and enriched for bound and unbound fractions using protein A magnetic beads. (**B**) Selected gene hits for regulation of EGFR degradation by Ctx-M6Pn. Hits with positive effect are in blue, and hits with negative effect are in red. (**C**) Schematic of hits categorized by subcellular localization or processes according to GO annotations, color coded by casTLE score. Genes with positive effect are in blue, and genes with negative effect are in red.

identified as hits with positive effect for Ctx-M6Pn–mediated degradation and were not enriched from the Ctx screen, which suggests that neddylation of CUL3 is important for LYTAC-mediated degradation (Fig. 4B). Knockout of genes (*CUL3*, *UBA3*, *CAND1*) in this pathway (fig. S11, B and C, and table S1) confirmed that degradation of EGFR was dependent on a functional E3 neddylation system (Fig. 4C and fig. S12A). To determine whether a CUL3 dependency was specific to EGFR degradation, we tested for degradation of c-MET in

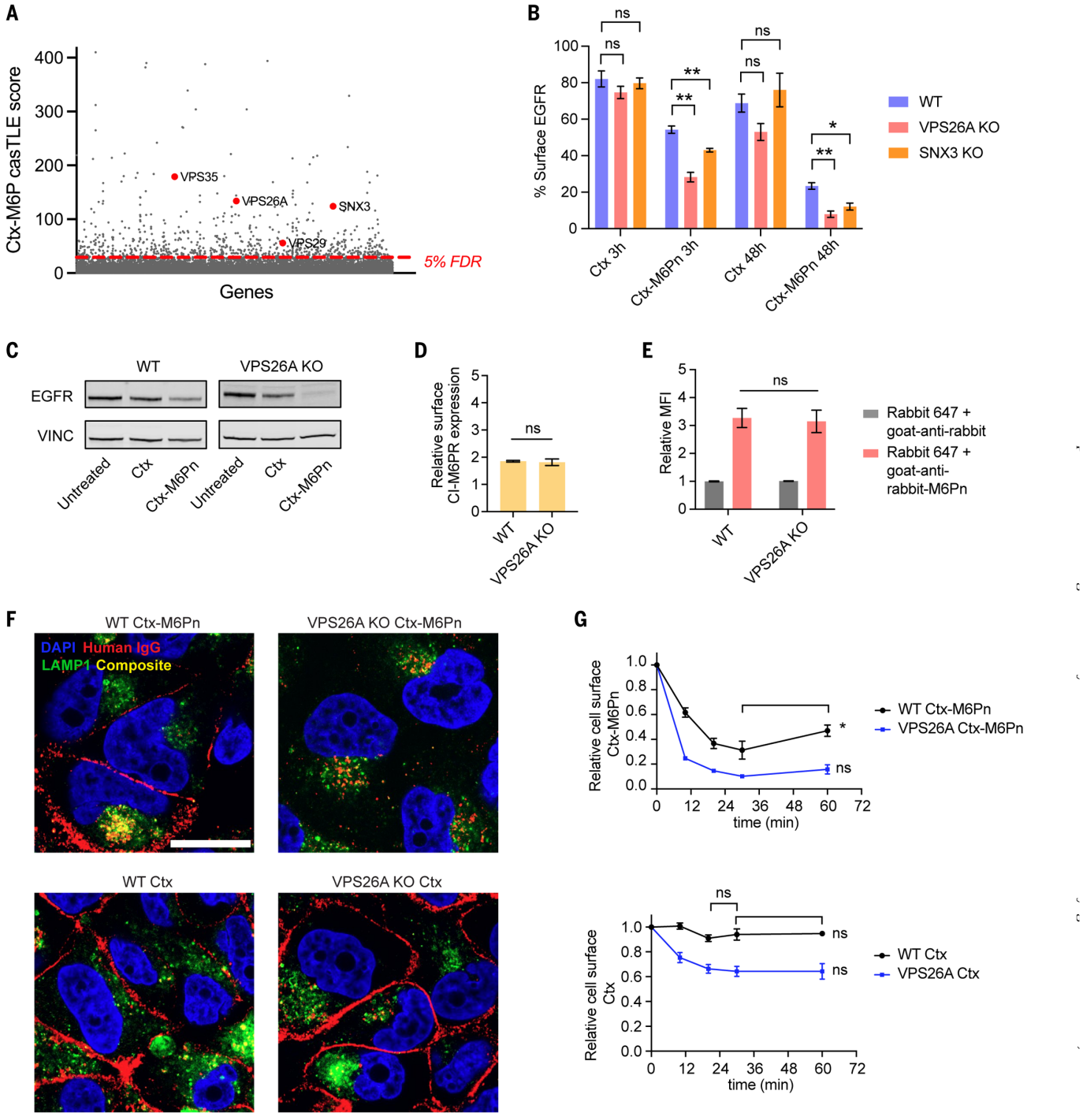


Fig. 3. Disrupting retromer complex genes enhances CI-M6PR–mediated target degradation by reducing the recycling of LYTAC–CI-M6PR complexes. (**A**) Retromer complex genes are negative hits from the CRISPR screen. (**B**) Degradation of EGFR in UMRC2 WT, VPS26A KO, and SNX3 KO cells as determined by live-cell flow cytometry after 3 hours or 48 hours of treatment with 10 nM Ctx or Ctx-M6Pn. (**C**) Immunoblot analysis of EGFR levels in WT or VPS26A KO cells after treatment with 10 nM Ctx or Ctx-M6Pn for 48 hours. (**D**) Cell surface expression level of CI-M6PR in WT and VPS26A KO cells by live-cell flow cytometry. (**E**) MFI relative to the control (rabbit IgG-647 only) for WT and VPS26A KO cells incubated at 37°C for 1 hour with 50 nM rabbit IgG-647 and 25 nM goat anti-rabbit or goat anti-rabbit M6Pn. MFI was determined by live-cell flow cytometry. (**F**) Localization of EGFR and Ctx-M6Pn or Ctx (goat anti-human-647) in WT and VPS26A KO cells after pulse treatment of Ctx-M6Pn. Cells were treated with 10 nM Ctx-M6Pn for 24 hours, then washed and incubated with fresh media for additional 24 hours. Scale bar, 10 μm. DAPI, 4′,6-diamidino-2-phenylindole. (**G**) Recycling of Ctx or Ctx-M6Pn determined by live-cell flow cytometry. UMRC2 WT and VPS26A KO cells were pulse treated with 10 nM Ctx or Ctx-M6Pn for 24 hours, then washed and incubated with fresh media for the indicated times followed by surface staining with anti-human-647 on ice. For (B), (D), (E), and (G), data represent three independent experiments and are shown as means ± SEMs. *P* values were determined by unpaired two-tailed *t* tests. NS, not significant; **P* ≤ 0.05; ***P* ≤ 0.01. Data in (C) are representative of three independent experiments. Images in (F) are representative of three independent experiments and are a single plane.

CUL3-deficient cells and found that c-MET degradation was also attenuated (Fig. 4D and fig. S12B), which suggests that the role of CUL3 is not target specific. The impact of neddylation was also examined with MLN4924, which effectively blocked neddylation of CUL3 and completely abrogated LYTAC-mediated degradation (fig. S13A). To verify this effect in a different cell line, we also confirmed that MLN4924 inhibited neddylation of CUL3 and prevented LYTAC-mediated degradation of EGFR and c-MET in HeLa cells. Unlike UMRC2 cells, HeLa cells down-regulate EGFR with EGF treatment. Treatment with MLN4924 had little effect on the EGF-mediated degradation of EGFR, which suggests that CUL3 neddylation is dispensable for EGFR down-regulation by endogenous mechanisms but is essential for LYTAC-mediated degradation (Fig. 4E and fig. S13, B and C). We also determined that recycling of neddylation is essential for LYTAC activity because treatment with a deneddylation inhibitor, CSN5i, abolished LYTAC activity (fig. S14). This is consistent with previous studies that have shown that dynamic neddylation and deneddylation are requirements for active cullin ligases (*37*, *38*).

To determine the stage of the LYTAC-mediated degradation pathway dependent on neddylation of CUL3, we first observed that the cell surface expression of CI-M6PR was unchanged in CUL3-deficient cells (Fig. 4F). We then asked whether CUL3 affects the internalization of LYTAC-target complexes. Although no reduction in total EGFR levels was observed after LYTAC treatment in CUL3-deficient cells (Fig. 4C), we observed a reduction in cell surface EGFR levels relative to the untreated control after LYTAC treatment (Fig. 4G). We confirmed that the binding and internalization of LYTACs did not change in CUL3-deficient cells (Fig. 4, H and I, and fig. S15). Loss of CUL3 also did not affect the LYTAC-mediated internalization of a soluble IgG (fig. S16). Thus, CUL3 activity is downstream of LYTAC-mediated internalization. To further investigate this effect, we looked at the localization of EGFR after 72 hours of LYTAC treatment and found that EGFR was concentrated partially in LE/LY vesicles in CUL3-deficient cells, whereas EGFR was mostly degraded in the WT cells (Fig. 4J). In addition, we performed a pulse-chase experiment where we treated with Ctx or LYTAC for 24 hours and then washed and incubated with fresh media. After 48 hours, LYTAC was cleared in WT cells, as indicated by the lack of anti-human-647 signal. However, LYTAC was observed in puncta that colocalized with CI-M6PR in CUL3-deficient cells (fig. S17), which suggests inefficient transport of LYTAC into lysosomes. Thus, CUL3 plays an essential role in LYTAC-mediated target maturation from endosome to lysosome for degradation.

To understand how CUL3 may regulate maturation of endosomes to lysosomes, we identified interactors of neddylated CUL3 through immunoprecipitation (IP)–proteomics. Because CUL3 is activated by neddylation, we looked for interactors that were depleted when neddylation was inhibited by MLN4924 treatment (Fig. 4K). Substrates that are ubiquitinated by CUL3 were identified by ubiquitin enrichment proteomics in WT and CUL3-deficient cells (Fig. 4L). A prominent hit was sequestosome-1 (SQSTM1) (Fig. 4, K and L). We confirmed enrichment of SQSTM1 after CUL3-IP in the absence of MLN4924 by immunoblot (fig. S18A). SQSTM1 is ubiquitinated by the E3 ligase RNF26, which allows its recruitment of vesicle adaptors involved in endosomal maturation and efficient cargo transfer (*39*). Our proteomics results suggest that ubiquitination of SQSTM1 by CUL3 is essential for late endosomal maturation and LYTAC activity. We confirmed that LYTACs had reduced ability to promote degradation of EGFR in SQSTM1 KO cells (fig. S18, B to D). Rather, EGFR was trapped in LE/LY compartments after LYTAC treatment (fig. S18E), similarly to CUL3 KO cells. Furthermore, SQSTM1 colocalized with Ctx-M6Pn in CUL3-deficient cells, whereas this behavior was less pronounced in WT cells (Fig. 4M and fig. S18F). Thus, SQSTM1 interacts with LYTAC-containing endosomes but requires ubiquitination at the late endosome by CUL3 to enable further maturation and trafficking of LYTAC-target complexes to lysosomes.

Given the critical role of CUL3 neddylation for LYTAC activity, we questioned whether the expression of neddylated CUL3 could be correlated with degradation efficacy. We compared the expression levels of neddylated CUL3 and EGFR degradation ability across 11 cell lines spanning eight different tissues and observed that cell lines with higher levels of neddylated CUL3 exhibited better degradation of EGFR (Fig. 4, N to P). This correlation suggests that neddylated CUL3 could serve as a predictive marker for LYTAC-mediated membrane protein degradation.

Disruption of M6P biosynthesis genes enhances LYTAC efficacy

Lysosomal hydrolases are synthesized in the ER and transported to the TGN, where they are modified with M6P. CI-M6PR in the TGN recognizes M6P-modified lysosomal proteins and traffics them to late endosomes for eventual lysosomal delivery (*40*–*42*). Our screen for LYTAC-mediated degradation identified several glycosyltransferases with negative effect sizes that are involved in the M6P glycoprotein biosynthesis pathway, including alpha-1,2-mannosyltransferase (ALG9), alpha-1,6-mannosyltransferase (ALG12), mannosyl-oligosaccharide glucosidase (MOGS), glucosidase II alpha subunit (GANAB), *N*-acetylglucosamine-1-phosphate transferase (GNPTAB), and *N*-acetylglucosamine-1-phosphodiester alpha-*N*-acetylglucosaminidase (NAGPA) (Fig. 5, A and B; fig. S19A; and table S1). Transmembrane protein 251 (TMEM251) or lysosomal enzyme trafficking factor (LYSET), which was recently characterized as a crucial protein for M6P modification and trafficking of lysosomal hydrolases (*43*, *44*), was also identified (Fig. 5B). Knockouts of ALG12 or GNPTAB resulted in greater lysotracker signal (fig. S19, B to D), consistent with previous accounts of enlarged lysosomes in GNPTAB KO cells (*45*). Ctx-M6Pn–mediated degradation was enhanced in ALG12- or GNPTAB-deficient cells after 3, 24, and 48 hours of treatment (Fig. 5C). After 48 hours of treatment, we observed >90% depletion of surface EGFR in these cells. By confocal microscopy, some EGFR remained on the cell surface in WT cells after 24 hours of LYTAC treatment, whereas EGFR was nearly depleted from the cell surface in ALG12- or GNPTAB-deficient cells (Fig. 5D). We did not detect any surface-bound EGFR after LYTAC treatment in these cells, indicating enhanced internalization of target-LYTAC–CI-M6PR complexes. We also examined c-MET and CA9 degradation in ALG12 and GNPTAB KO cells. After 3 hours of treatment with Ona-M6Pn, about 40% of surface c-Met was degraded in WT cells. However, we observed 70% degradation of c-Met in ALG12- and GNPTAB-deficient cells, which was the D_{max} (maximal percent degradation) reached after 48 hours in WT cells (Fig. 5E). Degradation of CA9 was similarly improved in GNPTAB-deficient cells with Gir-M6Pn, which confirms that disruption of M6P biosynthesis enhances LYTAC activity across multiple membrane targets (Fig. 5F).

M6P biosynthesis attenuates CI-M6PR cell surface accessibility

To elucidate how the disruption of M6P biosynthesis genes improves LYTAC efficacy, we first assessed internalization of soluble cargo at 37°C using rabbit IgG-647. We observed a significant increase in uptake in ALG12- and GNPTAB-deficient cells, where internalized cargo colocalized with lysotracker (Fig. 6, A and B). This may be attributed to increased expression of cell surface CI-M6PR. However, loss of ALG12 or GNPTAB did not result in significant differences in surface CI-M6PR expression (Fig. 6C). Reduced M6P biosynthesis may result in altered CI-M6PR occupancy by lysosomal hydrolases that traffic from the TGN to the plasma membrane, thereby increasing the fraction of unoccupied cell surface CI-M6PR accessible to LYTACs. To test this, we measured the binding of LYTAC to the cell surface by staining with rabbit IgG-647 and goat-anti-rabbit-M6Pn without internalization. Binding of a LYTAC to the cell surface was higher in

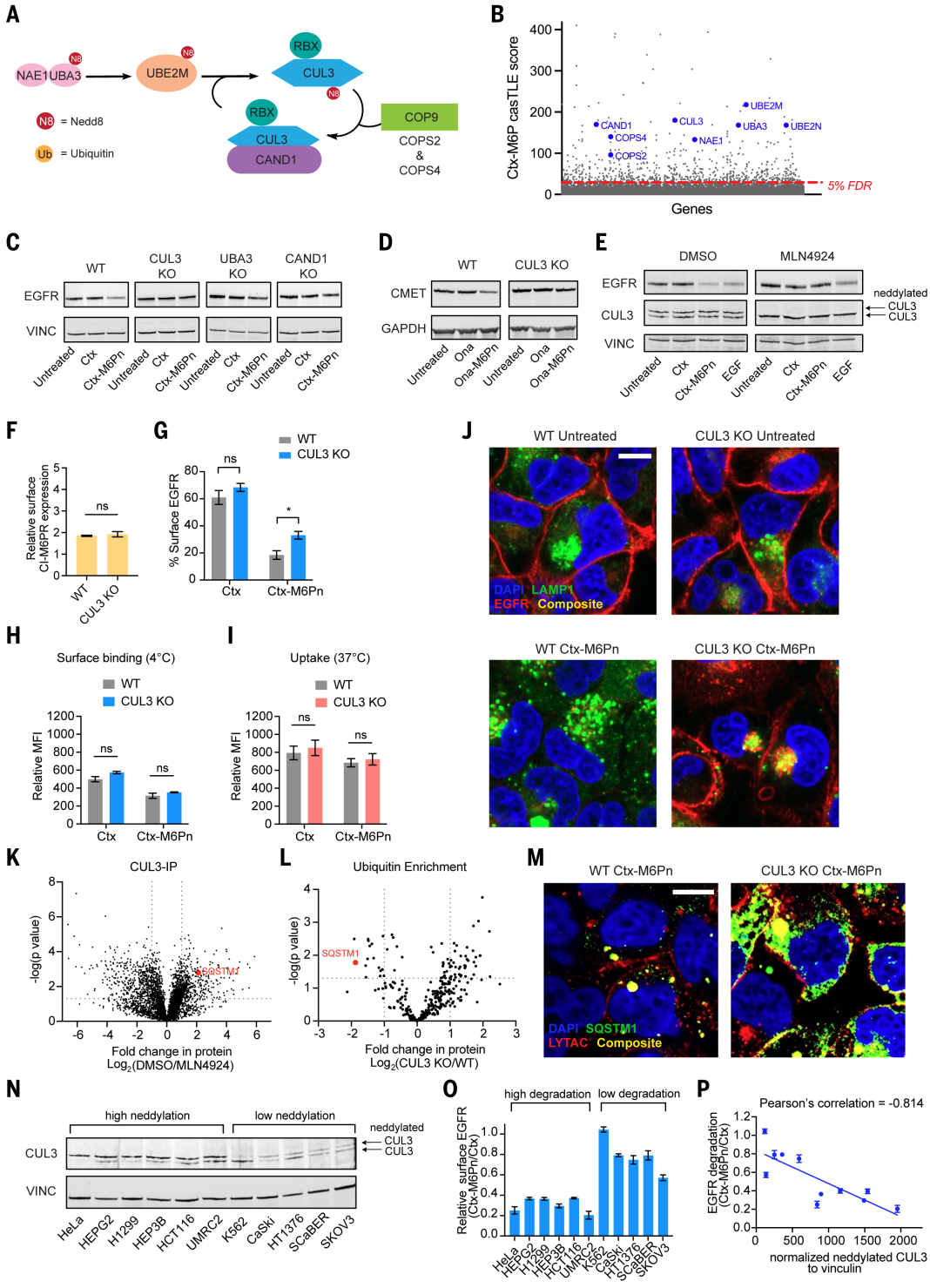


Fig. 4. Neddylation of CUL3 is essential for CI-M6PR–mediated target lysosomal degradation. (**A**) CUL3 neddylation process. Neddylation activates E3 ligase, CUL3. (**B**) Genes involved in CUL3 neddylation are positive hits from the CRISPR screen. (**C**) Immunoblot analysis of EGFR levels in WT and CUL3, UBA3, and CAND1 KO cells after treatment with 10 nM Ctx or Ctx-M6Pn for 48 hours. (**D**) Levels of c-MET in WT and CUL3 KO cells after treatment with 10 nM Ona or Ona-M6Pn for 48 hours. (**E**) Immunoblot analysis of EGFR levels in HeLa cells pretreated with DMSO or MLN4924 (2 μM) for 24 hours, then treated with 10 nM Ctx or Ctx-M6Pn for 24 hours. (**F**) Cell surface expression level of CI-M6PR in WT and CUL3 KO cells by live-cell flow cytometry. (**G**) Degradation of cell surface EGFR in WT and CUL3 KO cells as determined by live-cell flow cytometry after 48 hours of treatment with 10 nM Ctx or Ctx-M6Pn. (**H**) Surface staining of Ctx or Ctx-M6Pn with goat-anti-human-647 in WT or CUL3 KO cells measured by live-cell flow cytometry after 1.5-hour incubation on ice. (**I**) Live-cell MFI relative to the control (human IgG-647 only) for WT or CUL3 KO cells incubated at 37°C for 1.5 hours with 50 nM human IgG-647 and 25 nM Ctx or Ctx-M6Pn. (**J**) Visualization of EGFR degradation in WT and CUL3 KO cells using confocal microscopy after continuous treatment with Ctx-M6Pn for 72 hours. Scale bar, 10 μm. (**K**) Volcano plot of interactors after CUL3 immunoprecipitation in cells treated with or without MLN4924 (2 μM). (**L**) Volcano plot of ubiquitin enrichment via quantitative proteomics in WT and CUL3 KO cells. (**M**) Visualization of SQSTM1 and Ctx-M6Pn (IgG) in WT or CUL3 KO cells using confocal microscopy after 48-hour treatment with 10 nM Ctx-M6Pn. Scale bar, 10 μm. (**N**) Immunoblot analysis of CUL3 and neddylated CUL3 across various cell lines. (**O**) Degradation EGFR in various cell lines as determined by relative surface EGFR after 48-hour treatment with 10 nM Ctx-M6Pn versus Ctx by live-cell flow cytometry. (**P**) A scatter plot of EGFR degradation (Fig. 3K) versus normalized neddylated CUL3 expression to vinculin (Fig. 3L) with calculated Pearson's correlation. For (F), (G), (I), (J), and (O), data represent three independent experiments, and data are shown as means ± SEMs. *P* values were determined by unpaired two-tailed *t* tests. NS, not significant; *$P \leq 0.05$. For (C), (D), (E), (M), and (N), data represent three independent experiments. For (H), (J), (M), and (N), data are representative of two independent experiments. Images are a single plane.

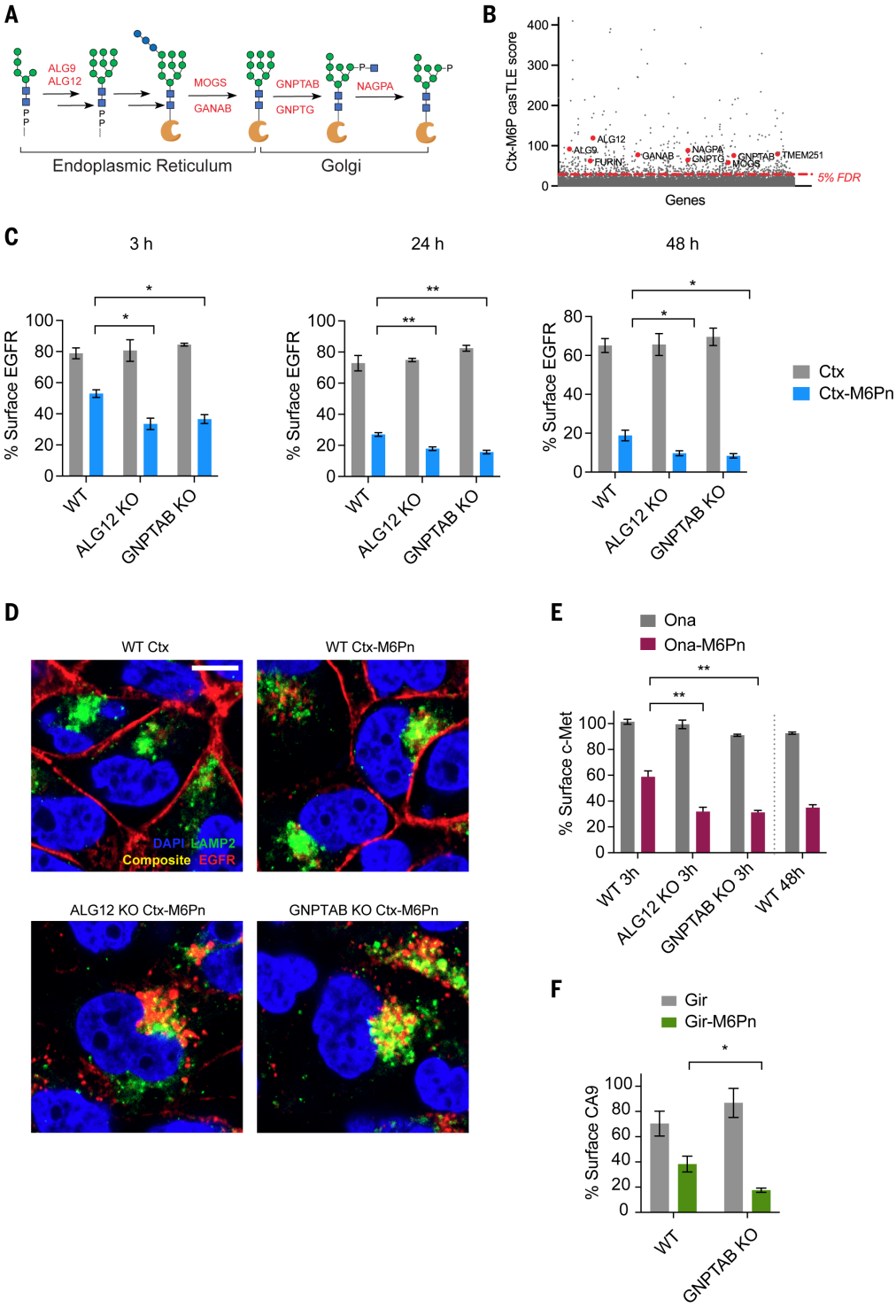


Fig. 5. Knockout of M6P biosynthesis genes enhances LYTAC efficacy. (**A**) M6P–N-glycan biosynthesis of lysosomal hydrolases in the ER and Golgi. (**B**) Genes involved in M6P biosynthesis pathway are negative hits from the CRISPR screen. (**C**) Degradation of EGFR in UMRC2 WT, ALG12 KO, and GNPTAB KO cells as determined by live-cell flow cytometry after 3 hours, 24 hours, and 48 hours of treatment with 10 nM Ctx or Ctx-M6Pn. (**D**) Visualization of EGFR degradation in WT and GNPTAB KO cells using confocal microscopy after continuous treatment with Ctx or Ctx-M6Pn for 24 hours. Scale bar, 10 μm. (**E**) Degradation of c-MET in UMRC2 WT, ALG12 KO, and GNPTAB KO cells as determined by live-cell flow cytometry after 3 hours and 48 hours of treatment with 10 nM Ctx or Ctx-M6Pn. (**F**) Degradation of CA-9 in UMRC2 WT and GNPTAB KO cells as determined by live-cell flow cytometry after 24 hours of treatment with 10 nM Ctx or Ctx-M6Pn. For (C), (E), and (F), data represent three independent experiments, and data are shown as means ± SEMs. *P* values were determined by unpaired two-tailed *t* tests. NS, not significant; $*P \leq 0.05$; $**P \leq 0.01$. Images in (D) are representative of two independent experiments and are a single plane.

ALG12- and GNPTAB-deficient cells and was ablated by cotreatment with inhibitory monomeric M6P (mM6P), which indicates that the binding is CI-M6PR dependent (Fig. 6D). Thus, a majority of cell surface CI-M6PR remains occupied at steady state.

We sought to identify proteins occupying cell surface CI-M6PR using quantitative surface proteomics comparing GNPTAB KO cells with WT cells, where we labeled the cell surface with cell-impermeable NHS-sulfo-biotin on ice and enriched cell lysates for biotinylated surface proteins (Fig. 6E). Quantitative proteomics showed a decrease in lysosomal proteins or hydrolases, such as mammalian ependymin-related protein 1 (EPDR1), N-acetylglucosamine-6-sulfatase (GNS), β-galactosidase (β-gal), and acid ceramidase (ASAH1), on the cell surface of GNPTAB-deficient cells compared with the WT cells (Fig. 6F). GNS (*46*, *47*), β-gal (*46*, *48*), and ASAH1 (*46*, *47*) are well-known lysosomal hydrolases that have M6P modifications. EDPR1 is a lysosomal protein with unknown function, but it is modified with M6P (*46*, *47*, *49*–*51*).

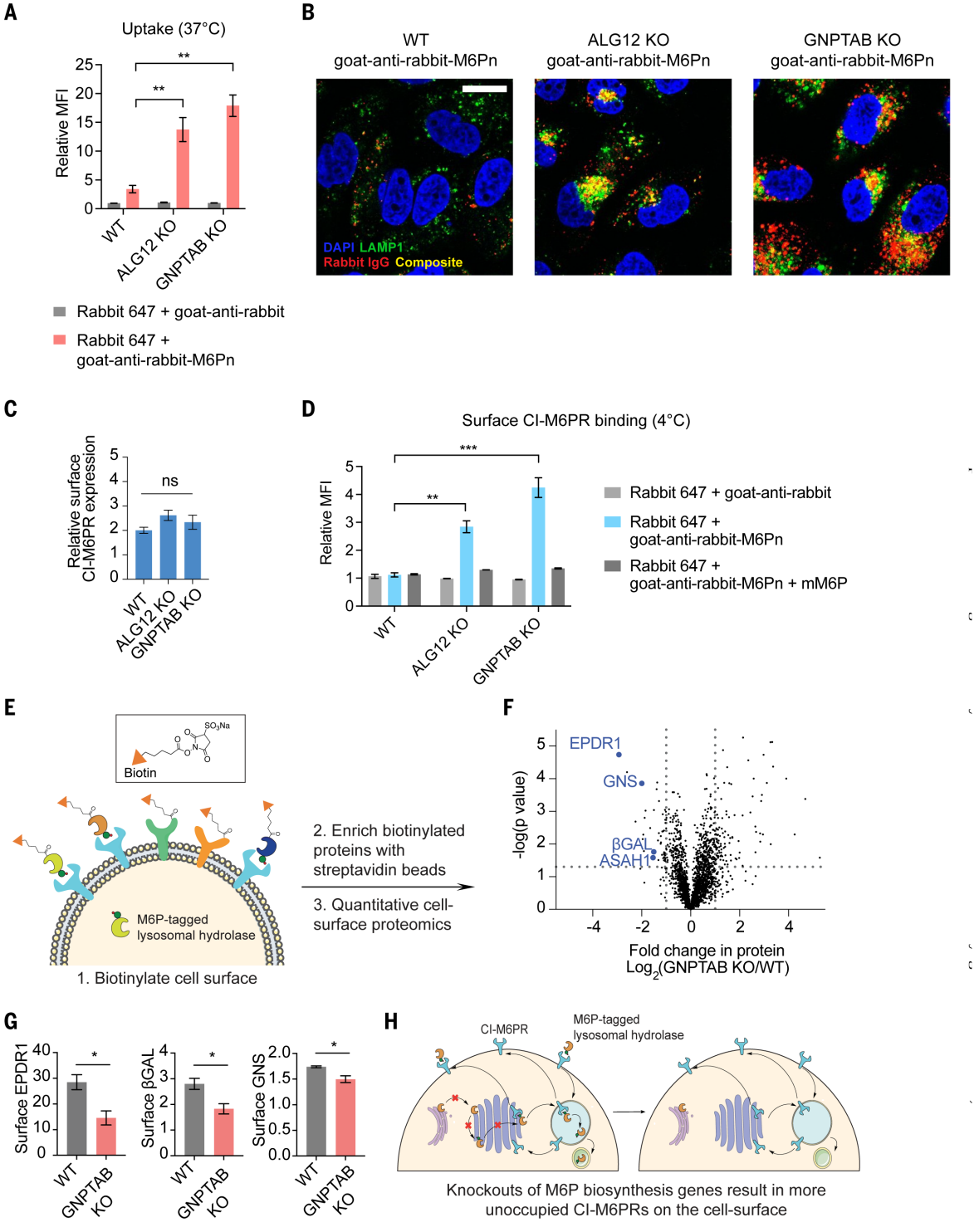


Fig. 6. M6P biosynthesis attenuates CI-M6PR cell surface accessibility. (**A**) Uptake of rabbit IgG-647. MFI relative to the control (rabbit IgG-647 only) for WT, ALG12 KO, and GNPTAB KO cells incubated at 37°C for 1 hour with 50 nM rabbit IgG-647 and 25 nM goat anti-rabbit or goat anti-rabbit M6Pn. MFI was determined by live-cell flow cytometry. (**B**) Live-cell imaging of UMRC2 WT, ALG12 KO, and GNPTAB KO cells that were incubated at 37°C for 1 hour with 50 nM rabbit IgG-647 and 25 nM goat anti-rabbit, goat anti-rabbit M6Pn. Scale bar, 20 μm. (**C**) Cell surface expression level of CI-M6PR in WT, ALG12 KO, and GNPTAB KO cells by live-cell flow cytometry. (**D**) Cell surface CI-M6PR binding. MFI relative to the control (rabbit IgG-647 only) for WT, ALG12 KO, and GNPTAB KO cells incubated at 4°C for 30 min with 50 nM rabbit IgG-647 and 25 nM goat anti-rabbit or goat anti-rabbit M6Pn. MFI was determined by live-cell flow cytometry. (**E**) Experimental setup for surface proteomics. WT or GNPTAB KO cells were incubated with NHS-sulfo-biotin on ice for 30 min and were enriched with streptavidin magnetic beads. (**F**) Relative quantitative surface proteomics between WT and GNPTAB KO cells. (**G**) Cell surface expression level of EPDR1, B-gal, and GNS in WT and GNPTAB KO cells. (**H**) Model for knockout of M6P biosynthesis genes resulting in an increased fraction of accessible CI-M6PR on the cell surface. For (A), (C), (D), and (G), data represent three independent experiments, and data are shown as means ± SEMs. *P* values were determined by unpaired two-tailed *t* tests. NS, not significant; **P* ≤ 0.05; ***P* ≤ 0.01; ****P* ≤ 0.001. Images in (B) are representative of two independent experiments and are a single plane. For (F), data are representative of three independent experiments.

We confirmed that levels of EPDR1 were also reduced in GNPTAB-deficient cells by immunoblot (fig. S18E), consistent with prior work demonstrating lower cellular levels of M6P-tagged proteins in GNPTAB KO cells (*45*). To validate our findings from surface proteomics, we analyzed the cell surface expression level of EPDR1, β-gal, and GNS by flow cytometry and confirmed that cell surface levels of these proteins were lower in the GNPTAB-deficient cells (Fig. 6G). Thus, CI-M6PR is occupied with M6P-modified proteins on the cell surface at

a steady state. As a result, perturbation of M6P biosynthesis genes increases the fraction of accessible CI-M6PRs on the cell surface, thereby enhancing internalization of LYTACs (Fig. 6H).

Discussion

First-generation LYTACs could not achieve beyond 70 to 80% target degradation after 24 to 48 hours of treatment, regardless of the specific protein target (*7*). These findings suggested that unknown aspects of the LYTAC endolysosomal trafficking mechanism may place a limit on protein degradation. We were curious whether this constraint could be overcome through a more nuanced understanding of these trafficking pathways. As with other degradation modalities, we observed that the D_{max} values differ across cell lines, even when the same protein is targeted. Discovery of features that drive this variation might allow the prediction of efficacy of LYTAC degradation through a priori genetic or biochemical analyses. In this work, we determined that disruption of retromer genes reduced the recycling of CI-M6PR–LYTAC complexes from endosomes to the plasma membrane and increased lysosomal degradation to >90%. From a therapeutic standpoint, knocking out retromer complexes may not be desirable because retromer complexes are important for conserved cellular endocytic processes (*52*). However, these results motivate antibody engineering of LYTACs to release their targets at low pH in endosomes, which could allow the target to degrade regardless of LYTAC recycling and could markedly enhance degradation efficacy. In addition, we identified that neddylation of CUL3 is an essential factor for late endosomal maturation and a predictive marker for LYTAC activity.

CI-M6PR–mediated lysosomal transport is the standard ERT approach for several lysosomal storage disorders. Yet, regulators of CI-M6PR internalization ability and trafficking are not fully understood. Recent deorphanization of TMEM251 as an essential factor for lysosomal enzyme transport and M6P trafficking machinery has highlighted the relevance of functional CI-M6PR–mediated transport in disease contexts (*43*, *44*). Through elucidating the mechanistic principles of LYTACs, we show that M6P biosynthesis attenuates CI-M6PR cell surface accessibility, and perturbation of M6P biosynthesis genes increases the fraction of unoccupied receptors on the cell surface. The observation of discrete M6P-tagged proteins that occupy receptors at the cell surface is supported by prior work which has demonstrated that free receptors were replaced with ligand-occupied receptors at the cell surface of cells treated with weak bases or monensin, which inhibit dissociation of receptor-ligand complexes (*53*). Although direct pharmacological targeting of M6P biosynthesis might adversely affect lysosomal enzyme activities, our findings reveal important design considerations for future LYTACs with enhanced efficacy. One strategy may be to use a higher-affinity M6P-ligand for CI-M6PR that can outcompete the lysosomal hydrolases. Alternatively, usage of orthogonal binding sites on CI-M6PR, such as the IGF2, urokinase receptor (uPAR), and retinoic acid binding sites, might prove more effective (*54*–*57*) (*41*). Beyond LYTACs, delivery applications dependent on CI-M6PR trafficking, including ERTs or macromolecule delivery platforms, could benefit from targeting alternative regions of the internalizing receptor. The guiding principles that result from our studies can be implemented in other therapeutics that benefit from enhanced lysosomal trafficking via CI-M6PR.

The field of extracellular protein degradation is emerging as a powerful method for therapeutic modulation of deleterious proteins. In this work, we used an unbiased functional genomic-driven mechanistic study to characterize the cellular determinants of an extracellular protein degradation platform (LYTACs). The genomic screen approach via protein A magnetic enrichment that we demonstrate can be applied to any other extracellular antibody-based degraders, such as PROTABs, KineTACs, and AbTACs, that harness different lysosome trafficking receptors and pathways. The unbiased screen resulted in fundamental information that sheds light on CI-M6PR occupancy and trafficking. Similar approaches may facilitate identification of the drivers of other degraders and the mechanistic nuances of alternative lysosome trafficking receptors that have been harnessed for degradation, such as ASGPR, RNF43, ZNRF3, and CXCR7 (*8*–*13*). Our work thus provides a clear roadmap for studying mechanistic principles of other extracellular degraders and a valuable lens for studying lysosome trafficking receptors.

Materials and methods

General chemical synthesis procedures

Reagent-grade chemical reagents were purchased from Carbosynth, Sigma Aldrich, Click Chemistry Tools, and TCI. All chemical reactions were performed in standard, flame-dried glassware capped with rubber septa under an inert atmosphere of nitrogen unless stated otherwise. Stainless steel syringes or cannulae were used to transfer moisture-sensitive liquids. Thin layer chromatography (TLC) was conducted on precoated glass plates covered with 0.2 mm silica gel for monitoring reactions. TLC plates were visualized with UV light or 5% H_2SO_4 in MeOH. Reaction mixtures were purified via column chromatography using Biotage SNAP KP-Sil or Ultra C18 cartridges (10 to 100 g) with a Biotage Isolera Prime ACI automated fraction collector. See the supplementary text (Chemical Synthesis Procedures and Characterization) for detailed synthetic protocols and characterization of new compounds.

Chemical analysis instrumentation

Proton nuclear magnetic resonance (^{1}H NMR) spectra were recorded on a Varian 400 spectrometer and proton-decoupled carbon-13 NMR (^{13}C {^{1}H} NMR) spectra were recorded on a Varian 500 or Varian 400 spectrometer at 25°C. Spectra were reported in parts per million (ppm) downfield of tetramethylsilane and are referenced to the residual resonances of the protium NMR solvent [CD_3OD: 3.31 (methanol)] and carbon NMR solvent [CD_3OD: 49.00 (methanol)]. MestReNova (v12.0.3) was used for all chemical NMR analysis. Data are reported as chemical shift, multiplicity (br, broad; s, singlet; d, doublet; t, triplet; q, quartet; quin, quintet; sept, septet; m, multiplet), coupling constants in Hertz (Hz) and integration. High-resolution mass spectrometric data were obtained on a Thermo Exploris 240 Orbitrap mass spectrometer by the Stanford University Mass Spectrometry (SUMS) core facility.

Cell lines

Adherent cells were cultured in T75 flasks or 15-cm plates at 37°C in a 5% CO_2 atmosphere. HeLa, HEPG2, HEP3B, HCT116, NCI-H1299, K562, CaSki, HT1376, SCaBER, and SKOV3 cells were obtained from ATCC. UMRC2 was a generous gift from Dr. Erinn Rankin at Stanford University. HeLa, HEPG2, HEP3B, HT1376, and SCaBER were cultured in Dulbecco's minimum essential medium (DMEM) supplemented with 10% heat-inactivated fetal bovine serum (HI-FBS) and 1% penicillin/streptomycin. NCI-H1299, K562, CaSKi were cultured in RPMI supplemented with 10% HI-FBS and 1% penicillin/streptomycin. HCT116 and SKOV3 were cultured in McCoy 5A supplemented with 10% HI-FBS and 1% penicillin/streptomycin.

Antibody-M6Pn conjugation

A 2 to 5 mg/ml solution of antibody (goat-anti-rabbit, Ctx, Gir) was buffer exchanged into phosphate-buffered saline (PBS) using a 7K Zeba size-exclusion column according to the manufacturer's protocol. To the antibody solution was added 50 equiv. of NHS-$(PEG)_4$-azide [20 mg ml^{-1} in dimethyl sulfoxide (DMSO), Click Chemistry Tools], and the reaction was incubated overnight at room temperature (RT). The reaction mixture was filtered using a 7K Zeba size-exclusion column to yield the conjugated antibody.

$M6Pn_5$ or $M65Pn_2$-BCN glycopeptides (100 equiv.) were weighed into an Eppendorf tube, and a 2 mg ml^{-1} solution of antibody-$(PEG)_4$-N_3 was added. The reaction was manually agitated until the mixture was homogeneous. The reaction mixture was allowed to incubate

at RT in the dark for 3 days and was filtered using a 7K Zeba size-exclusion column.

IgG internalization assay by flow cytometry

Cells were plated (75,000 cells per well in a 24-well plate) 1 day before the experiment. For rabbit-IgG internalization, cells were incubated with 250 μl of complete growth media with 50 nM Rabbit IgG-647 and 25 nM goat anti-rabbit, or goat anti-rabbit-M6Pn for 1 hour. For Ctx or LYTAC internalization, cells were incubated with 250 μl of complete growth media with 50 nM goat-anti-human-647 and 25 nM Ctx or 25 nM Ctx-M6Pn. After incubation, cells were washed with PBS three times, lifted with trypsin, then transferred to a 96-well V-bottom plate. The cells were washed three times with PBS + 0.5% bovine serum albumin (BSA) + 5 mM EDTA, then incubated with Sytox Green for 15 min on ice before flow cytometry analysis.

IgG internalization and colocalization by live-cell imaging

Cells were plated (50,000 cells per well in an 8-well Labtek) a day before the experiment. Cells were incubated with 200 μl of complete growth media with 50 nM Rabbit IgG-647 and 25 nM goat anti-rabbit or goat anti-rabbit-M6Pn for 1 hour. After incubation, 50 nM lysotracker was added and cells were incubated for additional 15 min at 37°C. Cells were then washed with PBS three times, incubated with Hoechst for 5 min, and imaged by confocal microscopy.

Protein degradation analysis by immunoblot

Cells were plated (75,000 cells per well in a 24-well plate) 1 day before the experiment. Cells were incubated with 250 μl of complete growth media with 10 nM LYTAC or controls for the indicated time. Cells were then washed with PBS three times and lysed with RIPA buffer supplemented with protease inhibitor cocktail (Roche), 0.1% Benzonase (Millipore-Sigma), and PhosStop (Roche) on ice for 30 min. The cell suspension was scraped and transferred to Eppendorf tubes, and spun down at 21,000 *g* for 15 min at 4°C. The supernatant was collected and the protein concentration was determined by BCA assay (Pierce). Equal amounts of lysates were loaded onto 4 to 12% Bis-Tris gel and separated by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE). Then, the gel was transferred onto a nitrocellulose membrane and stained with REVERT Total Protein Stain (LI-COR), then blocked with Intercept Blocking Buffer (TBS) (LI-COR) for 1 hour at RT. The membrane was incubated with primary antibody overnight at 4°C, washed three times with TBS-T. Subsequently, the membrane was incubated with secondary antibody for 1 hour at RT and washed three times with TBS-T for visualization with an Odyssey CLx Imager (LI-COR). Image Studio (LI-COR) was used to quantify band intensities.

Cell surface antigen staining by flow cytometry

Cells were lifted with trypsin for <5 min, quenched with media, and transferred to a 96-well V-bottom plate. Cells were washed three times with PBS + 0.5% BSA + 5 mM EDTA [fluorescence-activated cell sorting (FACS) buffer] and incubated with primary antibody for 30 min on ice. Cells were washed three times with PBS + 0.5% BSA + 5 mM EDTA (FACS buffer) and incubated with secondary antibody for 30 min on ice. After washing three times with FACS buffer, cells were incubated with either Sytox Green, Sytox Red, or Sytox Blue for 15 min on ice. Flow cytometry was performed on either a BD LSR II or MACSQuant Analyzer 10 Flow Cytometer, and FlowJo software was used to gate on single cells and live cells for analysis.

Cell surface degradation analysis by flow cytometry

Adherent cells were plated (75,000 cells per well in a 24-well plate) 1 day before the experiment. Cells were incubated with 250 μl of complete growth media with 10 nM LYTAC or controls for indicated time. Cells were then washed with PBS three times, lifted with trypsin for <5 min, quenched with media, and transferred to a 96-well V-bottom plate for staining by flow cytometry as shown above.

Quantitative proteomics for CA9 degradation

Sample processing

UMRC2 cells were plated (375,000 cells per well in a 6-well plate) 1 day before the experiment. Cells were incubated with 1500 μl of complete growth media with PBS (control), 10 nM Gir, or 10 nM Gir-M6Pn for 48 hours, with all conditions performed in triplicate. Cells were then washed with PBS and lysed with RIPA buffer supplemented with protease inhibitor cocktail (Roche), 0.1% Benzonase (Millipore-Sigma), and phosphatase inhibitor cocktail (Cell Signaling Technologies) on ice for 30 min. The cells were scraped, transferred to Eppendorf tubes, and spun down at 21,000 *g* for 15 min at 4°C. The supernatant was collected, protein concentration was determined using BCA assay (Pierce), and ~30 μg of total protein from each sample was processed for quantitative bottom-up proteomics using an micro S-trap (ProtiFi) protocol as follows. The concentration of SDS was brought to 5% SDS and disulfide bonds were reduced by heating samples to 95°C for 10 min after addition of 5 mM DTT. After cooling down to RT for 3 min, reduced cysteine residues were alkylated by incubating in 30 mM iodoacetamide (Sigma, A3221) in the dark at RT for 30 min. Samples were then acidified with phosphoric acid at a final concentration of 1.2% and vortexed. Bind/wash buffer [100 mM tetraethylammonium bromide (TEAB) in 90% methanol] was then added to each sample (seven times the final volume after acidification). 150 μl of each sample were loaded onto micro S-trap columns and spun down at 4000 *g* for 20 s. Flow-through was discarded, and the centrifugation step was repeated until all the solution passed through the column. After three washes with 150 μl bind/wash buffer, trypsin (Promega, V5113) in 50 mM TEAB solution was added to the S-trap at a 1:25 protease:protein ratio and incubated at 47°C for 90 min. After trypsinization, peptides were eluted from the S-trap with 40 μl 50 mM TEAB, 40 μl 0.2% formic acid, and 40 μl 50% acetonitrile and 0.2% formic acid by spinning down at 1000 *g* for 60 s for each elution step. All elution steps were collected in the same tube for a give sample and the total peptide mixtures were lyophilized. Peptides were resuspended in 0.2% formic acid, normalized to concentration using a Nanodrop Spectrophotomerter (Thermo Fisher, absorbance at 205 nm), and analyzed by liquid chromatography–tandem mass spectrometry (LC-MS/MS). Before analysis, one microliter of each sample was taken and combined into a pooled sampled that was used to make the chromatogram library discussed below.

LC-MS/MS data acquisition

Proteomics data were acquired using a spectrum-library free data-independent acquisition (DIA) approach that relies on gas-phase fractionation (GPF) to generate DIA-only chromatogram libraries (*58*, *59*). Peptides were separated over a 25 cm Aurora Series Gen2 reverse-phase LC column (75 μm inner diameter packed with 1.6 μm FSC C18 particles, Ion Opticks). The mobile phases (A: water with 0.2% formic acid; B: acetonitrile with 0.2% formic acid) were driven and controlled by a Dionex Ultimate 3000 RPLC nano system (Thermo Fisher). An integrated loading pump was used to load peptides onto a trap column (Acclaim PepMap 100 C18, 5 μm particles, 20 mm length, Thermo Fisher) at 5 μl/min, which was put in line with the analytical column 5.5 min into the gradient. The gradient was held at 0% B for the first 6 min of the analysis, followed by an increase from 0% to 5% B from 6 to 6.5 min, and increase from 5 to 22% B from 6.5 to 66.5 min, an increase from 22% to 90% from 66.5 to 71 min, isocratic flow at 90% B from 71 to 75 min, and re-equilibration at 0% B for 15 min for a total analysis time of 90 min per acquisition. Eluted peptides were analyzed on an Orbitrap Fusion Tribrid MS system (Thermo Fisher). Precursors were ionized was ionized with a spray voltage held at +2.2 kV relative to ground, the RF lens was set to 60%, and the inlet capillary temperature was held at 275°C.

Six chromatogram library files were collected through six repeated injections of the

pooled sample only. The instrument was configured to acquire 4 m/z precursor isolation window DIA spectra using a staggered isolation window pattern (*60*) from narrow mass ranges using window placements optimized by Skyline. DIA MS/MS spectra were acquired with an AGC target of 400,000 charges, a maximum injection time of 54 ms, beam-type collisional dissociation (i.e., HCD) with a normalized collision energy of 33, and a resolution of 30,000 at 200 m/z using the Orbitrap as a mass analyzer. The six gas-phase fractionation chromatogram libraries were collected with nominal mass ranges of 400-500 m/z, 500-600 m/z, 600-700 m/z, 700-800 m/z, 800-900 m/z, and 900-1000 m/z. The exact windowing scheme was downloaded from https://bitbucket.org/searleb/encyclopedia/wiki/Home (*58*) and is available in the supplementary materials. Precursor MS1 spectra were interspersed every 25 scans with an AGC target of 400,000 charges, a maximum injection time of 55 ms, a resolution of 60,000 at 200 m/z using the Orbitrap as a mass analyzer, and a scan range of either 395-505 m/z, 495-605 m/z, 595-705 m/z, 695-805 m/z, 795-905 m/z, or 895-1005 m/z. For quantitative samples (i.e., the non-pooled samples) the instrument was configured to acquire 25 × 16 m/z precursor isolation window DIA spectra covering 385-1015 m/z using a staggered isolation window pattern with window placements optimized by Skyline (windowing scheme downloaded from the same link as above and available in the supplementary materials). DIA spectra were acquired with the same MS/MS settings described above. Precursor MS1 spectra were interspersed every 38 scans with a scan range of 385-1015 m/z, an AGC target of 400,000 charges, a maximum injection time of 55 ms, and a resolution of 60,000 at 200 m/z using the Orbitrap as a mass analyzer. The detailed parameters were recorded (file included in the supplementary materials). To ensure chromatographic conditions were consistent between library generation and quantitative sample acquisition, GPF library runs were acquired in the middle of the set of quantitative samples described below, and the quantitative samples were collected in randomized order.

Data analysis

Staggered DIA spectra were demultiplexed from raw data into mzML files with 10 ppm accuracy using MSConvert (*61*) with settings described in Pino *et al.* (*58*). Encyclopedia (version 1.12.31) (*59*) was used to search demultiplexed mzML files using an internal PECAN fasta search engine called Walnut (*62*) and a reviewed human proteome database (canonical sequences only) from Uniprot (*63*). Walnut settings were: fixed cysteine carbamidomethylation, full tryptic digestion with up to 2 missed cleavages, HCD (y-only) fragmentation, 10 ppm precursor and fragment mass tolerances, and 5 quantitative ions. The chromatogram library resulting from the Walnut search was then used for Encyclopedia searching, where all analogous settings to the Walnut search remained the same, and other settings included a library mass tolerance of 10 ppm, inclusion of both b- and y-type fragment ions, and a minimum number of quantitative ions set at 3. Percolator (version 3.1) was used to filter peptides to a 1% false discovery rate (FDR) using the target/decoy approach and proteins to a 1% protein-level FDR assuming protein grouping parsimony. Resulting data from EncyclopeDIA were checked in Skyline (*64*) before further processing with Perseus (*65*), including filtering so that only proteins that were detected in all replicates of all three conditions were retained and significance calculations for comparisons of Gir and Gir-M6Pn treatments relative to control were performed using a two-tailed *t* test with a permutation-based FDR with 250 randomizations, an FDR of 0.05, and an S0 value of 1.

Genome-wide CRISPR-Cas9 screen in UMRC2 cells

UMRC2 cells were transduced with LentiArray Cas9 Lentivirus (Invitrogen A32064) and single-cell sorted to obtain stable UMRC2-Cas9 cells. A genome-wide, 10 sgRNA per gene CRISPR deletion library from the Bassik laboratory at Stanford University (*29*) was synthesized, cloned, and infected into Cas9-expressing UMRC2 cells. Briefly, ~400 million UMRC2-Cas9 cells (4 million cells per 15-cm plate) were infected with the CRISPR KO library at a multiplicity of infection (MOI) of 0.3 to 0.4. Cells expressing sgRNAs were selected for using puromycin (12 μg/ml) for 5 days. Selected cells were then plated for treatments in puromycin-free media.

For the screen, cells were split into two conditions, each in duplicate: Ctx (10 nM) treated group and Ctx-M6Pn (10 nM) treated group. For each replicate and condition, 100 million cells were plated and treated with Ctx or Ctx-M6Pn for 48 hours. Cells were washed with PBS and harvested (~200 million cells per replicate). Cells were resuspended in FACS buffer (0.5% BSA + 2 mM EDTA in PBS) then split into 15 ml falcon tubes with maximum 10 ml per tube (4 million cells/ml). Pierce Protein A Dynabeads were added to the cells (15 μl per 2 million cells) and incubated with rotation at 4°C for 30 min. After incubation, the tubes were placed on a magnetic rack for 2 min. The unbound supernatant was transferred to a new tube (unbound fraction), placed on the magnetic rack again for 2 min to remove any remaining beads, and the supernatant was pelleted and saved. The beads were washed twice with FACS buffer, and the remaining beads were saved and frozen as the bound fraction.

Genomic DNA of each condition was extracted using Qiagen DNA Blood Maxi kit (catalog no. 51194). The sgRNA sequences were amplified and prepared for sequencing as previously described (*29*). These libraries were then sequenced using an Illumina NextSeq with ~40 million reads per condition. Analysis and comparison of guide composition of Ctx-M6Pn treated versus Ctx treated conditions were performed using casTLE as previously described (*30*).

Generation and analysis of knockout cell lines

For all genes except *CUL3* and *SNX3*, UMRC2 cells were nucleofected with a Lonza 4D-Nucleofector Unit using the Lonza SE Cell Line 4D-Nucleofector Kit. sgRNAs and spCas9 were obtained from Synthego's Gene Knockout Kit v2, and nucleofection was carried out using manufacturer's protocol. For *CUL3* or *SNX3*, sgRNA targeting the gene was cloned into the lentiCRISPR v2 vector (Addgene, no. 52961) and lentivirus was generated in HEK293T cells and transduced into UMRC2 cells. Cells were subsequently selected with puromycin (12 μg/ml) for two passages.

DNA was extracted using QuickExtract DNA Extraction Solution (Epicentre, catalog no. QE09050) and polymerase chain reaction (PCR) amplification was performed according to Synthego's "Genotyping" protocol. Knockout analysis was performed using Synthego's online tool, Inference of CRISPR Edits (ICE), according to the manufacturer's protocol.

sgRNA CUL3: GAATGATCATCAAACAGCTA

sgRNA UBA3: ACGUUUCUCACUUCAGUGCU, GCUCGAGGAACUUCUUUACA, GCAGAAUGGCUGUUGAUGGU

sgRNA CAND1: GUGAUGAUGAUGACAUGAGU, UGGAUGCUGUAGUUAGCACA, AGAGCGUGAAGAGAAUGUAA

sgRNA VPS26A: UUAUUUCAGAGUUUUCUUGG, UGAUGGGGAAACCAGGAAAA, CGGAGAAUCCGUUUCAGGAA

sgRNA SNX3: GGGGTCCGTAGGCGTCATTC

sgRNA ALG12: AGCGUAGCUCCUUGUGUGGC, GACGCACGCGCCGACGGUGC, GCAGCCCAGGCCGCGGGGCA

sgRNA GNPTAB: AGUGACAAUAGUAACACACC, AUGAAAAUAUUCCGAACCCA, UUGAAGAUAACGAAGAACUG

sgRNA SQSTM1: UAUGGCGUCGCUCACCGUGA, CUGCAGCCCCGAGCCUGAGG, CUGCGAGCGGCUGCUGAGCC

Confocal microscopy for membrane protein degradation

Adherent cells were plated (50,000 cells per well in an 8-well chamber slide) 1 day before the experiment. Cells were incubated with 200 μl of complete growth media with 10 nM LYTAC or controls for indicated time. Cells were then washed with PBS and fixed with 4% paraformaldehyde in PBS for 15 min at RT,

washed three times, and permeabilized with 0.1% Triton for 5 min at RT. Cells were blocked in 10% goat serum in PBS for 1 hour at RT, and incubated with primary antibody overnight at 4°C. Cells were washed with DPBS, then incubated with secondary antibody and DAPI for 1 hour at RT. Cells were washed with DPBS and imaged with Nikon A1R confocal microscope using Plan Fluor 60x oil immersion 1.30-numerical aperture objective. 405-nm violet laser, 488-nm blue laser, 561-nm green laser, and 639-nm red laser were used. Single focal plane images were captured and are represented in figures.

Neddylation inhibition of CUL3

Adherent cells were plated (50,000 to 75,000 cells per well in a 24-well plate) 1 day before the experiment. Cells were pretreated with 2 μM MLN4924 for 24 hours. On the following day, cells were treated with fresh 2 μM MLN4924 and 10 nM Ctx or Ctx-M6Pn for another 24 hours. Cells were then harvested for immunoblot analysis as mentioned above.

Cell surface binding analysis by flow cytometry

Cells were trypsinized, quenched with media, and transferred to a 96-well V-bottom plate. Cells were washed three times with PBS + 0.5% BSA + 5 mM EDTA (FACS buffer) and incubated with 50 nM Rabbit IgG-647 and 25 nM goat anti-rabbit or goat anti-rabbit-M6Pn for 30 min on ice. Cells were washed three times with PBS + 0.5% BSA + 5 mM EDTA (FACS buffer) and incubated with Sytox Green. Flow cytometry was performed on either a BD LSR II or MACSQuant Analyzer 10 Flow Cytometer, and FlowJo software was used to gate on single cells and live cells for analysis.

CUL3-IP proteomics

UMRC2 cells were plated 1 day before the experiment (375,000 cells per well in a 6-well plate, 3 wells were used per condition per replicate). Cells were pretreated with 2 μM MLN4924 for 24 hours. On the following day, cells were treated with fresh 2 μM MLN4924 and 10 nM Ctx or Ctx-M6Pn for another 24 hours. Cells were then washed with PBS three times and lysed with RIPA buffer supplemented with protease inhibitor cocktail (Roche), 0.1% Benzonase (Millipore-Sigma), and PhosStop (Roche) on ice for 30 min (100 μl of lysis buffer per 1 well of 6-well plate). Lysates were spun down at 21,000 *g* for 15 min at 4°C. The supernatant was collected and the protein concentration was determined by BCA assay (Pierce). Pierce Protein A Dynabeads were precomplexed with rabbit anti-CUL3 antibody (Novus) (4 μl of antibody in 40 μl of Dynabeads per sample) for 2 hours rotating at RT. Equal amounts of lysates were incubated with antibody-bead complex rotating at 4°C overnight. The next day, the beads were washed three times with RIPA lysis buffer and eluted by boiling at 95°C for 10 min with 5% SDS with 20 mM DTT. The quantitative proteomics workflow for CUL3-IP proteomics was similar to the methods used to investigate CA9 degradation as described above. Samples were digested with trypsin using the micro S-trap protocol, peptide abundance was measured with a Nanodrop Spectrophotomerter (absorbance at 205 nm), and one microliter of each sample combined into a pooled sample that was used to make a chromatogram library for CUL3-IP samples. The same DIA approach and LC-MS/MS data acquisition strategy was used as described above. Data analysis was similar, but instead of EncyclopeDIA, demultiplexed mzML files were processed with FragPipe using the MSFragger-DIA and DIA-NN (*66*, *67*) for quantitative analysis with default settings. Data were further processed with Perseus as described above: protein intensity values were $\log_2$ transformed before further analysis, proteins with measured values for all three replicates with at least one condition were retained, and missing values were imputed from a normal distribution with a width of 0.3 and a downshift value of 2.5. Significance calculations for pairwise comparisons were performed using a two-tailed *t* test with a permutation-based FDR with 250 randomizations, an FDR of 0.05, and an S0 value of 1.

Ubiquitin-enrichment proteomics

Sample preparation

UMRC2 cells were plated 1 day before the experiment (15-cm plate). The cells were treated with 10 nM Ctx or Ctx-M6Pn for 48 hours. Cells were then washed with PBS three times and lysed with RIPA buffer supplemented with protease inhibitor cocktail (Roche), 0.1% Benzonase (Millipore-Sigma), and PhosStop (Roche) on ice for 30 min (100 μl of lysis buffer per 1 well of 6-well plate). Lysates were spun down at 21,000 *g* for 15 min at 4°C. The supernatant was collected, and the protein concentration was determined by BCA assay (Pierce).

Samples were processed with an S-trap protocol as described above, except modified for the midi columns to digest 5 mg of lysate (as measured by BCA) per condition. After digestion, 10 μg of peptides were kept for quantitative proteomics with DIA methods described above, whereas the rest of the peptides were processed for ubiquitin site profiling with TMT labeling using methods adapted from previous work (*68*–*70*). The PTM-Scan ubiquitin remnant motif (K-ε-GG) kit (Cell Signaling Technology, kit no. 5562) was used to enrich for the ubiquitin remnant motif. For all steps in the enrichment of K-ε-GG peptides except peptide elution, microcentrifuge tubes containing the sample and the antibody beads are kept on ice when not spinning. First, antibody-bound beads were washed three times with 100 mM sodium borate (pH 9.0) and incubated with 20 mM DMP for 60 min at RT. Beads were washed twice with 200 mM ethanolamine and incubated overnight at 4°C with 200 mM ethanolamine. After this incubation, beads were washed three times with immunoprecipitation (IAP) buffer (50 mM MOPS, pH 7.2, 10 mM sodium phosphate, 50 mM NaCl). For each enrichment, 156 μg of crosslinked anti-K-ε-GG antibody beads was used. Peptides from each sample were resuspended in 1 mL of ice-cold IAP buffer and vortexed. The pH of each sample was checked by spotting ~1 μL of solution on pH paper to ensure the pH was ~7. Peptide samples were then centrifuged at 20,000 *g* for 5 min to remove any insoluble material. Peptides were added to a tube of aliquoted antibody and incubated for 2 hours at 4°C, with gentle end-over-end rotation. After incubation, all samples were centrifuged at 2000 *g* for 1 min and then antibody beads were allowed to settle by letting tubes sit for ~10 to 20 s on ice. The supernatant (IP flowthrough) was removed. This IP flowthrough was immediately transferred to another aliquot of crosslinked anti-K-ε-GG antibody beads for a second enrichment, repeated as described above, meaning each lysate was immunoprecipitated twice. During the 2-hour incubation of the second IP, antibody beads from the first IP were washed with all washes completed as quickly as possible. For washing the beads, after the addition of the wash reagent, the tubes were inverted about five times, centrifuged for 30 s at 2000 *g*, allowed to sit for 10 to 20 s to let the beads settle, and the supernatant was removed. Immediately after removal of the IP flowthrough, two washes with 1 mL ice cold IAP buffer were followed by three washes with 1 mL ice-cold PBS. DiGly peptides were eluted twice with 50 μL 0.15% TFA, by first gently tapping the tube after addition of TFA and letting them set at RT for 5 min. Beads with TFA were then centrifuged at 2000 *g* for 1 min, beads were allowed to settle for 10 s, and supernatants containing the eluted peptides were removed and pooled. This washing and elution process was repeated for the second IP, and elutions from the first and second IP were combined before desalting with Strata-X reversed phase cartridges and drying with vacuum centrifugation. After enrichment, diGly peptides were labeled with TMTpro 16-plexkit (*71*) according to manufacturer's instructions (Thermo Fisher Scientific). TMT reagents were equilibrated to RT and resuspended in 20 μL anhydrous ACN. DiGly peptides were resuspended in 20 μL 100 mM TEAB and combined with appropriate TMT labels for a 1-hour incubation at RT. After this, 5 μL of 5% hydroxylamine was added to each sample with a 15-min incubation at RT. TMT-labeled samples were then combined and dried with vacuum centrifugation. Combined

samples were then fractionated with a high pH reversed phase spin column kit according to manufacturer's instructions (Thermo Fisher Scientific) with slight modifications. Six fractions were collected with elution buffers containing 17.5%, 20%, 22.5%, 25%, 30%, and 70% ACN, respectively, made with 0.1% triethylamine as the aqueous fraction. Spin columns were washed once with 300 µL ACN and twice with 0.1% TFA by spinning at 5000 *g* for 2 min. TMT-labeled diGly peptides were resuspended in 0.1% TFA and added to the column with centrifugation for 2 min. The columns were then washed with 300 µL 5% ACN in 0.1% TFA and 300 µL water before collecting six fractions using the elution buffers with increasing ACN percentage described above. Fractions were dried with vacuum centrifugation and resuspended in 0.2% formic acid for LC-MS/MS analysis. TMT labeling conditions are included in table S2.

LC-MS/MS data acquisition

Nonlabeled, nonenriched peptides from these samples were processed with the same DIA methods described above. TMT-labeled diGly peptides were separated over a 25 cm Aurora Series Gen2 reverse-phase LC column (75 µm inner diameter packed with 1.6 µm FSC C18 particles, Ion Opticks). The mobile phases (A: water with 0.2% formic acid; B: acetonitrile with 0.2% formic acid) were driven and controlled by a Dionex Ultimate 3000 RPLC nano system (Thermo Fisher). An integrated loading pump was used to load peptides onto a trap column (Acclaim PepMap 100 C18, 5 µm particles, 20 mm length, Thermo Fisher) at 5 µl/min, which was put in line with the analytical column 5.5 min into the gradient. The gradient was held at 0% B for the first 6 min of the analysis, followed by an increase from 0% to 5% B from 6 to 6.5 min, and increase from 5 to 25% B from 6.5 to 200 min, an increase from 25% to 90% from 200 to 218 min, isocratic flow at 90% B from 218 to 224 min, and re-equilibration at 0% B for 16 min for a total analysis time of 240 min per acquisition. Eluted peptides were analyzed on an Orbitrap Fusion Tribrid MS system (Thermo Fisher). Precursors were ionized was ionized with a spray voltage held at +2.2 kV relative to ground, the RF lens was set to 60%, and the inlet capillary temperature was held at 275°C. Survey scans of peptide precursors were collected in the Orbitrap from 400 to 1600 Th with an AGC target of 400,000, a maximum injection time of 50 ms, and a resolution of 120,000 at 200 m/z. Monoisotopic precursor selection was enabled for peptide isotopic distributions, precursors of z = 2 to 6 were selected for data-dependent MS/MS scans for 3 s of cycle time, and dynamic exclusion was set to 120 s with a ±10 ppm window set around the precursor monoisotope. An isolation window of 0.7 m/z was used to select precursor ions with the quadrupole, and peptides were analyzed using a synchronous precursor selection (SPS) MS3 method (*72*). Precursors were fragmented by CID at a normalized collision energy (NCE) of 35%. After acquisition of each MS2 spectrum, an SPS-MS3 scan was collected on the top 10 most intense ions in the MS2 spectrum. SPS-MS3 precursors were fragmented by high-energy collision–induced dissociation (HCD) and analyzed using the Orbitrap (NCE = 50%, AGC = 100,000, maximum injection time = 118 ms, and resolution = 60K).

Data analysis

Whole proteome DIA samples were analyzed as described above. Data from the TMT diGly peptides were searched using searched with the Andromeda search engine in MaxQuant (*73*, *74*) using the entire human proteome downloaded from Uniprot9 (reviewed, 20,428 entries). Quantitation settings for "Reporter ion MS3" with TMT-16plex reporter ion masses were used with a reporter ion mass tolerance set to 0.003. Cleavage specificity was set to Trypsin/P with 3 missed cleavages, and variable modifications included the diGly residue on lysine, oxidation of methionine, and acetylation of the protein N terminus with five maximum modifications per peptide. Defaults were used for the remaining settings, including PSM and protein FDR thresholds of 0.01 and 20 ppm, 4.5 ppm, and 0.5 Da for first search MS1 tolerance, main search MS1 tolerance, and MS2 product ion tolerance, respectively. Quantified diGly sites were then processed in Perseus. Contaminants and reverse hits were removed, results were filtered for diGly that had localization probabilities > 0.75, and signal in all relevant TMT channels was required. Significance testing was performed using a two-tailed *t* test with 250 randomizations and an FDR of 0.05.

Quantitative cell surface proteomics

UMRC2 WT cells or GNPTAB-deficient cells were plated in triplicate in 15-cm plates (10 million cells each) 1 day before the experiment. Cells were washed with cold PBS and incubated in PBS (no biotin control) or 1 mg/ml solution of NHS-sulfo-biotin in PBS with gentle rocking for 30 min at 4°C. Biotin labeling was quenched with 100 mM Tris-HCl, and the cells were washed three times with cold PBS. Cells were scraped using cold PBS and were pelleted and lysed with RIPA buffer supplemented with protease inhibitor cocktail (Roche), 0.1% Benzonase (Millipore-Sigma), and phosphatase inhibitor cocktail (Cell Signaling Technologies) on ice for 30 min. Cells were then clarified by centrifugation at 21,000 *g* for 15 min at 4°C. The supernatant was collected, and the protein concentration was determined by BCA assay (Pierce). Equal amounts of lysates were then incubated with 500 µl of Streptavidin Dynabeads (Pierce) and rotated overnight at 4°C. For enrichment, the beads were pelleted using a magnetic rack and the supernatant was discarded. Beads were washed twice with cold RIPA lysis buffer, once with 1M KCl, once with 0.1M sodium bicarbonate, once with 2M urea in 10 mM Tris-HCl pH 8.0, then twice with RIPA lysis buffer. Biotinylated proteins were eluted from the beads by boiling each sample in 5% SDS supplemented with 2 mM biotin and 20 mM DTT for 10 min. The beads were pelleted, and the eluates were collected and processed using the same S-trap protocol and DIA methods described above. For quantitative comparisons of cell surface protein difference, protein intensity values were $\log_2$ transformed before further analysis in Perseus, proteins with measured values for all three replicates with at least one condition were retained, and missing values were imputed from a normal distribution with a width of 0.3 and a downshift value of 1.8. Significance calculations for pairwise comparisons were performed using a two-tailed *t* test with a permutation-based FDR with 250 randomizations, an FDR of 0.05, and an S0 value of 1. GO keywords and cellular component terms were used to annotate proteins as lysosomal and cell surface.

Figure illustration

Parts of Fig. 1 and Fig. 6 were modified from Servier Medical Art, licensed under a Creative Common Attribution 3.0 Generic License.

REFERENCES AND NOTES

1. K. M. Sakamoto *et al*., Protacs: Chimeric molecules that target proteins to the Skp1–Cullin–F box complex for ubiquitination and degradation. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 8554–8559 (2001). doi: 10.1073/pnas.141230798; pmid: 11438690
2. G. E. Winter *et al*., Phthalimide conjugation as a strategy for in vivo target protein degradation. *Science* **348**, 1376–1381 (2015). doi: 10.1126/science.aab1433; pmid: 25999370
3. J. Krönke *et al*., Lenalidomide causes selective degradation of IKZF1 and IKZF3 in multiple myeloma cells. *Science* **343**, 301–305 (2014). doi: 10.1126/science.1244851; pmid: 24292625
4. G. Lu *et al*., The myeloma drug lenalidomide promotes the cereblon-dependent destruction of Ikaros proteins. *Science* **343**, 305–309 (2014). doi: 10.1126/science.1244917; pmid: 24292623
5. A. Mullard, First targeted protein degrader hits the clinic. *Nat. Rev. Drug Discov.* **18**, 237 (2019). doi: 10.1038/d41573-019-00043-6; pmid: 30936511
6. A. Mullard, Targeted protein degraders crowd into the clinic. *Nat. Rev. Drug Discov.* **20**, 247–250 (2021). doi: 10.1038/d41573-021-00052-4; pmid: 33737725
7. S. M. Banik *et al*., Lysosome-targeting chimaeras for degradation of extracellular proteins. *Nature* **584**, 291–297 (2020). doi: 10.1038/s41586-020-2545-9; pmid: 32728216
8. G. Ahn *et al*., LYTACs that engage the asialoglycoprotein receptor for targeted protein degradation. *Nat. Chem. Biol.* **17**, 937–946 (2021). doi: 10.1038/s41589-021-00770-1; pmid: 33767387
9. D. F. Caianiello *et al*., Bifunctional small molecules that mediate the degradation of extracellular proteins. *Nat. Chem. Biol.* **17**, 947–953 (2021). doi: 10.1038/s41589-021-00851-1; pmid: 34413525
10. A. D. Cotton, D. P. Nguyen, J. A. Gramespacher, I. B. Seiple, J. A. Wells, Development of Antibody-Based PROTACs for the Degradation of the Cell-Surface Immune Checkpoint Protein PD-L1. *J. Am. Chem. Soc.* **143**, 593–598 (2021). doi: 10.1021/jacs.0c10008; pmid: 33395526

11. H. Marei *et al*., Antibody targeting of E3 ubiquitin ligases for receptor degradation. *Nature* **610**, 182–189 (2022). doi: 10.1038/s41586-022-05235-6; pmid: 36131013
12. K. Pance *et al*., Modular cytokine receptor-targeting chimeras for targeted degradation of cell surface and extracellular proteins. *Nat. Biotechnol.* **41**, 273–281 (2023). doi: 10.1038/s41587-022-01456-2; pmid: 36138170
13. Y. Zhou, P. Teng, N. T. Montgomery, X. Li, W. Tang, Development of Triantennary N-Acetylgalactosamine Conjugates as Degraders for Extracellular Proteins. *ACS Cent. Sci.* **7**, 499–506 (2021). doi: 10.1021/acscentsci.1c00146; pmid: 33791431
14. K. A. Donovan *et al*., Mapping the Degradable Kinome Provides a Resource for Expedited Degrader Development. *Cell* **183**, 1714–1731.e10 (2020). doi: 10.1016/j.cell.2020.10.038; pmid: 33275901
15. M. Słabicki *et al*., The CDK inhibitor CR8 acts as a molecular glue degrader that depletes cyclin K. *Nature* **585**, 293–297 (2020). doi: 10.1038/s41586-020-2374-x; pmid: 32494016
16. C. Mayor-Ruiz *et al*., Plasticity of the Cullin-RING Ligase Repertoire Shapes Sensitivity to Ligand-Induced Protein Degradation. *Mol. Cell* **75**, 849–858.e8 (2019). doi: 10.1016/j.molcel.2019.07.013; pmid: 31442425
17. R. J. Desnick, E. H. Schuchman, Enzyme replacement and enhancement therapies: Lessons from lysosomal disorders. *Nat. Rev. Genet.* **3**, 954–966 (2002). doi: 10.1038/nrg963; pmid: 12459725
18. M. Gary-Bobo, P. Nirdé, A. Jeanjean, A. Morère, M. Garcia, Mannose 6-phosphate receptor targeting and its applications in human diseases. *Curr. Med. Chem.* **14**, 2945–2953 (2007). doi: 10.2174/092986707782794005; pmid: 18220730
19. J. Z. Drago, S. Modi, S. Chandarlapaty, Unlocking the potential of antibody-drug conjugates for cancer therapy. *Nat. Rev. Clin. Oncol.* **18**, 327–344 (2021). doi: 10.1038/s41571-021-00470-8; pmid: 33558752
20. M. Ritchie, L. Tchistiakova, N. Scott, Implications of receptor-mediated endocytosis and intracellular trafficking dynamics in the development of antibody drug conjugates. *MAbs* **5**, 13–21 (2013). doi: 10.4161/mabs.22854; pmid: 23221464
21. R. Kanasty, J. R. Dorkin, A. Vegas, D. Anderson, Delivery materials for siRNA therapeutics. *Nat. Mater.* **12**, 967–977 (2013). doi: 10.1038/nmat3765; pmid: 24150415
22. T. P. Prakash *et al*., Targeted delivery of antisense oligonucleotides to hepatocytes using triantennary *N*-acetyl galactosamine improves potency 10-fold in mice. *Nucleic Acids Res.* **42**, 8796–8807 (2014). doi: 10.1093/nar/gku531; pmid: 24992960
23. A. D. Springer, S. F. Dowdy, GalNAc-siRNA Conjugates: Leading the Way for Delivery of RNAi Therapeutics. *Nucleic Acid Ther.* **28**, 109–118 (2018). doi: 10.1089/nat.2018.0736; pmid: 29792572
24. K. R. Kampen, Membrane proteins: The key players of a cancer cell. *J. Membr. Biol.* **242**, 69–74 (2011). doi: 10.1007/s00232-011-9381-7; pmid: 21732009
25. G. Von Heijne, The membrane protein universe: What's out there and why bother? *J. Intern. Med.* **261**, 543–557 (2007). doi: 10.1111/j.1365-2796.2007.01792.x; pmid: 17547710
26. M. S. Haney *et al*., Identification of phagocytosis regulators using magnetic genome-wide CRISPR screens. *Nat. Genet.* **50**, 1716–1727 (2018). doi: 10.1038/s41588-018-0254-1; pmid: 30397336
27. B. Mair *et al*., High-throughput genome-wide phenotypic screening via immunomagnetic cell sorting. *Nat. Biomed. Eng.* **3**, 796–805 (2019). doi: 10.1038/s41551-019-0454-8; pmid: 31548591
28. R. A. Kamber *et al*., Inter-cellular CRISPR screens reveal regulators of cancer cell phagocytosis. *Nature* **597**, 549–554 (2021). doi: 10.1038/s41586-021-03879-4; pmid: 34497417
29. D. W. Morgens *et al*., Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. *Nat. Commun.* **8**, 15178 (2017). doi: 10.1038/ncomms15178; pmid: 28474669
30. D. W. Morgens, R. M. Deans, A. Li, M. C. Bassik, Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. *Nat. Biotechnol.* **34**, 634–636 (2016). doi: 10.1038/nbt.3567; pmid: 27159373
31. R. Rojas *et al*., Regulation of retromer recruitment to endosomes by sequential action of Rab5 and Rab7. *J. Cell Biol.* **183**, 513–526 (2008). doi: 10.1083/jcb.200804048; pmid: 18981234
32. M. Lucas *et al*., Structural Mechanism for Cargo Recognition by the Retromer Complex. *Cell* **167**, 1623–1635.e14 (2016). doi: 10.1016/j.cell.2016.10.056; pmid: 27889239
33. C. N. Arighi, L. M. Hartnell, R. C. Aguilar, C. R. Haft, J. S. Bonifacino, Role of the mammalian retromer in sorting of the cation-independent mannose 6-phosphate receptor. *J. Cell Biol.* **165**, 123–133 (2004). doi: 10.1083/jcb.200312055; pmid: 15078903
34. E. J. Bennett, J. Rush, S. P. Gygi, J. W. Harper, Dynamics of cullin-RING ubiquitin ligase network revealed by systematic quantitative proteomics. *Cell* **143**, 951–965 (2010). doi: 10.1016/j.cell.2010.11.017; pmid: 21145461
35. J.-T. Wu, H.-C. Lin, Y.-C. Hu, C.-T. Chien, Neddylation and deneddylation regulate Cul1 and Cul3 protein accumulation. *Nat. Cell Biol.* **7**, 1014–1020 (2005). doi: 10.1038/ncb1301; pmid: 16127432
36. Z.-Q. Pan, A. Kentsis, D. C. Dias, K. Yamoah, K. Wu, Nedd8 on cullin: Building an expressway to protein destruction. *Oncogene* **23**, 1985–1997 (2004). doi: 10.1038/sj.onc.1207414; pmid: 15021886
37. T. A. Soucy, P. G. Smith, M. Rolfe, Targeting NEDD8-activated cullin-RING ligases for the treatment of cancer. *Clin. Cancer Res.* **15**, 3912–3916 (2009). doi: 10.1158/1078-0432.CCR-09-0343; pmid: 19509147
38. S. Lyapina *et al*., Promotion of NEDD-CUL1 conjugate cleavage by COP9 signalosome. *Science* **292**, 1382–1385 (2001). doi: 10.1126/science.1059780; pmid: 11337588
39. M. L. M. Jongsma *et al*., An ER-Associated Pathway Defines Endosomal Architecture for Controlled Cargo Transport. *Cell* **166**, 152–166 (2016). doi: 10.1016/j.cell.2016.05.078; pmid: 27368102
40. S. Kornfeld, Lysosomal enzyme targeting. *Biochem. Soc. Trans.* **18**, 367–374 (1990). doi: 10.1042/bst0180367; pmid: 2164980
41. P. Ghosh, N. M. Dahms, S. Kornfeld, Mannose 6-phosphate receptors: New twists in the tale. *Nat. Rev. Mol. Cell Biol.* **4**, 202–212 (2003). doi: 10.1038/nrm1050; pmid: 12612639
42. J. R. Duncan, S. Kornfeld, Intracellular movement of two mannose 6-phosphate receptors: Return to the Golgi apparatus. *J. Cell Biol.* **106**, 617–628 (1988). doi: 10.1083/jcb.106.3.617; pmid: 2964450
43. C. Pechincha *et al*., Lysosomal enzyme trafficking factor LYSET enables nutritional usage of extracellular proteins. *Science* **378**, eabn5637 (2022). doi: 10.1126/science.abn5637; pmid: 36074822
44. C. M. Richards *et al*., The human disease gene LYSET is essential for lysosomal enzyme transport and viral infection. *Science* **378**, eabn5648 (2022). doi: 10.1126/science.abn5648; pmid: 36074821
45. E. van Meel *et al*., Multiple Domains of GlcNAc-1-phosphotransferase Mediate Recognition of Lysosomal Enzymes. *J. Biol. Chem.* **291**, 8295–8307 (2016). doi: 10.1074/jbc.M116.714568; pmid: 26833567
46. D. E. Sleat *et al*., Identification and validation of mannose 6-phosphate glycoproteins in human plasma reveal a wide range of lysosomal and non-lysosomal proteins. *Mol. Cell. Proteomics* **5**, 1942–1956 (2006). doi: 10.1074/mcp.M600030-MCP200; pmid: 16709564
47. D. E. Sleat, H. Zheng, P. Lobel, The human urine mannose 6-phosphate glycoproteome. *Biochim. Biophys. Acta* **1774**, 368–372 (2007). doi: 10.1016/j.bbapap.2006.12.004; pmid: 17258946
48. J. Distler, V. Hieber, G. Sahagian, R. Schmickel, G. W. Jourdian, Identification of mannose 6-phosphate in glycoproteins that inhibit the assimilation of β-galactosidase by fibroblasts. *Proc. Natl. Acad. Sci. U.S.A.* **76**, 4235–4239 (1979). doi: 10.1073/pnas.76.9.4235; pmid: 116230
49. Y. Wei *et al*., Crystal structures of human lysosomal EPDR1 reveal homology with the superfamily of bacterial lipoprotein transporters. *Commun. Biol.* **2**, 52 (2019). doi: 10.1038/s42003-018-0262-9; pmid: 30729188
50. D. E. Sleat *et al*., The human brain mannose 6-phosphate glycoproteome: A complex mixture composed of multiple isoforms of many soluble lysosomal proteins. *Proteomics* **5**, 1520–1532 (2005). doi: 10.1002/pmic.200401054; pmid: 15789345
51. D. E. Sleat, H. Zheng, M. Qian, P. Lobel, Identification of sites of mannose 6-phosphorylation on lysosomal proteins. *Mol. Cell. Proteomics* **5**, 686–701 (2006). doi: 10.1074/mcp.M500343-MCP200; pmid: 16399764
52. C. Burd, P. J. Cullen, Retromer: A master conductor of endosome sorting. *Cold Spring Harb. Perspect. Biol.* **6**, a016774 (2014). doi: 10.1101/cshperspect.a016774; pmid: 24492709
53. T. Braulke, C. Gartung, A. Hasilik, K. von Figura, Is movement of mannose 6-phosphate-specific receptor triggered by binding of lysosomal enzymes? *J. Cell Biol.* **104**, 1735–1742 (1987). doi: 10.1083/jcb.104.6.1735; pmid: 2953746
54. L. J. Olson *et al*., Allosteric regulation of lysosomal enzyme recognition by the cation-independent mannose 6-phosphate receptor. *Commun. Biol.* **3**, 498 (2020). doi: 10.1101/2020.02.21.959957
55. D. O. Morgan *et al*., Insulin-like growth factor II receptor as a multifunctional binding protein. *Nature* **329**, 301–307 (1987). doi: 10.1038/329301a0; pmid: 2957598
56. A. Nykjaer *et al*., Mannose 6-phosphate/insulin-like growth factor-II receptor targets the urokinase receptor to lysosomes via a novel binding interaction. *J. Cell Biol.* **141**, 815–828 (1998). doi: 10.1083/jcb.141.3.815; pmid: 9566979
57. J. X. Kang, J. Bell, A. Leaf, R. L. Beard, R. A. S. Chandraratna, Retinoic acid alters the intracellular trafficking of the mannose-6-phosphate/insulin-like growth factor II receptor and lysosomal enzymes. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 13687–13691 (1998). doi: 10.1073/pnas.95.23.13687; pmid: 9811861
58. L. K. Pino, S. C. Just, M. J. MacCoss, B. C. Searle, Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. *Mol. Cell. Proteomics* **19**, 1088–1103 (2020). doi: 10.1074/mcp.P119.001913; pmid: 32312845
59. B. C. Searle *et al*., Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat. Commun.* **9**, 5128 (2018). doi: 10.1038/s41467-018-07454-w; pmid: 30510204
60. D. Amodei *et al*., Improving Precursor Selectivity in Data-Independent Acquisition Using Overlapping Windows. *J. Am. Soc. Mass Spectrom.* **30**, 669–684 (2019). doi: 10.1007/s13361-018-2122-8; pmid: 30671891
61. R. Adusumilli, P. Mallick, Data Conversion with ProteoWizard msConvert. *Methods Mol. Biol.* **1550**, 339–368 (2017). doi: 10.1007/978-1-4939-6747-6_23; pmid: 28188540
62. Y. S. Ting *et al*., PECAN: Library-free peptide detection for data-independent acquisition tandem mass spectrometry data. *Nat. Methods* **14**, 903–908 (2017). doi: 10.1038/nmeth.4390; pmid: 28783153
63. UniProt Consortium, UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021). doi: 10.1093/nar/gkaa1100; pmid: 33237286
64. L. K. Pino *et al*., The Skyline ecosystem: Informatics for quantitative mass spectrometry proteomics. *Mass Spectrom. Rev.* **39**, 229–244 (2020). doi: 10.1002/mas.21540; pmid: 28691345
65. S. Tyanova *et al*., The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016). doi: 10.1038/nmeth.3901; pmid: 27348712
66. F. Yu *et al*., One-stop analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *bioRxiv*2022.10.28.514272 [Preprint] (2022); https://doi.org/10.1101/2022.10.28.514272.
67. V. Demichev *et al*., dia-PASEF data analysis using FragPipe and DIA-NN for deep proteomics of low sample amounts. *Nat. Commun.* **13**, 3944 (2022). doi: 10.1038/s41467-022-31492-0; pmid: 35803928
68. C. M. Rose *et al*., Highly Multiplexed Quantitative Mass Spectrometry Analysis of Ubiquitylomes. *Cell Syst.* **3**, 395–403.e4 (2016). doi: 10.1016/j.cels.2016.08.009; pmid: 27667366
69. N. D. Udeshi, P. Mertins, T. Svinkina, S. A. Carr, Large-scale identification of ubiquitination sites by mass spectrometry. *Nat. Protoc.* **8**, 1950–1960 (2013). doi: 10.1038/nprot.2013.120; pmid: 24051958
70. N. D. Udeshi *et al*., Rapid and deep-scale ubiquitylation profiling for biology and translational research. *Nat. Commun.* **11**, 359 (2020). doi: 10.1038/s41467-019-14175-1; pmid: 31953384
71. J. Li *et al*., TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing. *J. Proteome Res.* **20**, 2964–2972 (2021). doi: 10.1021/acs.jproteome.1c00168; pmid: 33900084
72. G. C. McAlister *et al*., MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Anal. Chem.* **86**, 7150–7158 (2014). doi: 10.1021/ac502040v; pmid: 24927332
73. J. Cox *et al*., Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011). doi: 10.1021/pr101065j; pmid: 21254760
74. S. Tyanova, T. Temu, J. Cox, The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016). doi: 10.1038/nprot.2016.136; pmid: 27809316

ACKNOWLEDGMENTS

We thank M. A. Gray (Stanford University) for providing Girentuximab, M. Abu-Remaileh (Stanford University) for helpful discussions, and T. McLaughlin and Stanford University Mass Spectrometry for high-resolution mass spectrometry characterization. **Funding:** This work was supported, in part, by a National Institutes of Health (NIH) grant R01GM058867 (C.R.B.). G.A. was supported by a National Science Foundation Graduate Research Fellowship and the Stanford Center for Molecular Analysis and Design. N.M.R. was supported by an NIH K99 Pathway to Independence Award (K99 GM1473043). S.W. is supported by funding from the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), the Cancer Research Society, and the Canadian Glycomics Network (GlycoNet). R.A.K. was supported by an NIH/NCI K99 Pathway to Independence Award (1K99CA259218-01). S.M.B. was supported by the Burroughs Wellcome Fund Careers at the Scientific Interface Award. **Author contributions:** G.A., S.M.B., and C.R.B. conceived the project. G.A., R.A.K., S.M.v.H., and N.M.R. carried out experiments and interpreted data. S.W. and M.C.B. oversaw and provided insights and materials. G.A., S.M.B., and C.R.B. wrote the manuscript with input from all authors. C.R.B. and S.M.B. provided supervision. **Competing interests:** Stanford University has filed patent applications relating to LYTACs, which are licensed to Lycia Therapeutics, listing G.A., S.M.B., and C.R.B. as coinventors. S.M.B. is a Scientific Advisory Board Member of Lycia Therapeutics. C.R.B. is a cofounder and Scientific Advisory Board member of Lycia Therapeutics, Palleon Pharmaceuticals, Enable Bioscience, Redwood Biosciences (a subsidiary of Catalent), and InterVenn Biosciences. The authors declare no other competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials. In addition, the mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD037899. Inquiries for materials in this manuscript can be directed to the corresponding authors. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse. This article is subject to HHMI's Open Access to Publications policy. HHMI lab heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the Author Accepted Manuscript (AAM) of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adf6249
Supplementary Text
Figs. S1 to 19
Tables S1 to S3
MDAR Reproducibility Checklist
Data S1 to S8

Submitted 3 November 2022; resubmitted 23 May 2023
Accepted 31 August 2023
10.1126/science.adf6249

RESEARCH ARTICLE SUMMARY

STRATOSPHERIC OZONE

Rapid ozone depletion after humidification of the stratosphere by the Hunga Tonga Eruption

Stephanie Evan*, Jerome Brioude, Karen H. Rosenlof, Ru-Shan Gao, Robert W. Portmann, Yunqian Zhu, Rainer Volkamer, Christopher F. Lee, Jean-Marc Metzger, Kevin Lamy, Paul Walter, Sergio L. Alvarez, James H. Flynn, Elizabeth Asher, Michael Todt, Sean M. Davis, Troy Thornberry, Holger Vömel, Frank G. Wienhold, Ryan M. Stauffer, Luis Millán, Michelle L. Santee, Lucien Froidevaux, William G. Read

INTRODUCTION: The Hunga Tonga–Hunga Ha'apai (HT) underwater volcanic eruption of 15 January 2022 was exceptional in that it injected more water (H_2O) vapor than has ever been observed in the satellite era. A rapid response was initiated that provided the data needed to explore the early chemical effects of this rare event. Volcanic eruptions inject gases and particles into the stratosphere, potentially influencing climate and ozone (O_3) chemistry. Understanding the intricate connections between volcanic emissions and atmospheric chemistry and their broader implications is vital for advancing our capabilities of modeling environmental processes.

RATIONALE: Our research provides new insights into the initial effects of the HT eruption on stratospheric composition. This eruption was notable for injecting material, including an unprecedented amount of H_2O vapor (10% of the total global mean stratospheric burden), to very high altitudes (up to 55 km). Using observational data, we investigated the speed of impacts and potential influence on stratospheric bromine and chlorine species, nitrogen oxide (NO), as well as O_3.

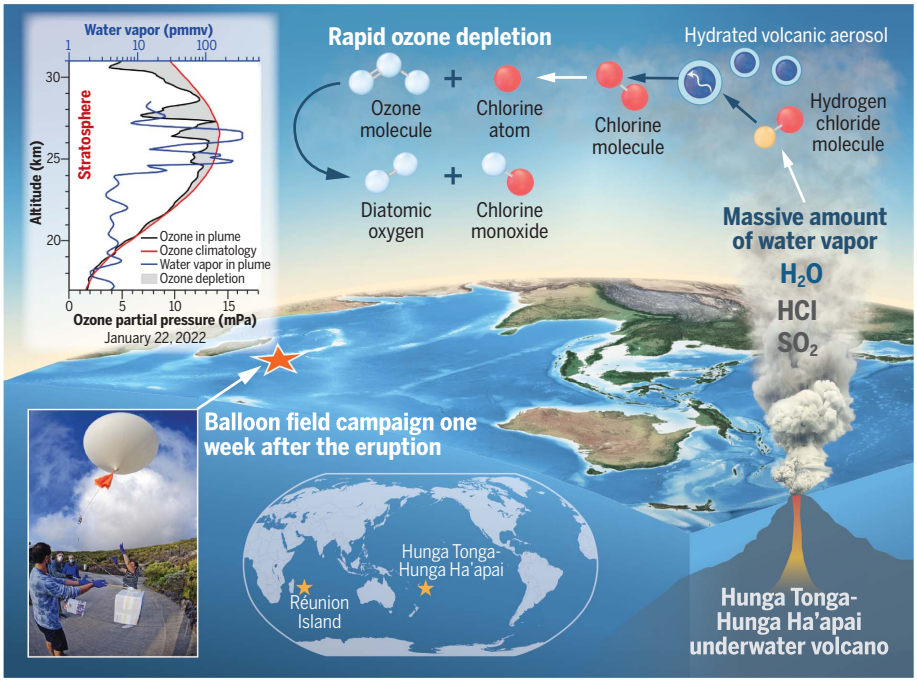
RESULTS: Our study combined balloon measurements, zenith sky observations, and satellite data to capture the early evolution of the HT volcanic plume's impact on O_3. In just 1 week, the stratospheric O_3 above the tropical southwestern Pacific and Indian Ocean region decreased by 5%. This change gains context when compared with the Antarctic O_3 hole, where up to 60% of the O_3 is depl... from September to November. This trop... O_3 decrease exceeds that of previous eruptions, underscoring the HT event's exceptional nature.

H_2O vapor played a pivotal role in consequences of the HT event. Its increased abundance resulted in higher relative humidity and radiative cooling in the stratosphere, enabling chemical reactions on the surfaces of volcanic aerosols to occur at temperatures warmer than their typical range. These chemical processes on hydrated volcanic aerosols led to the formation of active chlorine species such as chlorine monoxide (ClO) from inactive chlorine (hydrogen chloride, HCl). Combined with a decrease in nitrogen oxides (NOx), these active chlorine species catalytically destroyed O_3 molecules. The decrease in hydrogen chloride (HCl) by 0.4 parts per billion by volume (ppbv) and the increase in ClO by 0.4 ppbv provided compelling evidence for chlorine activation, leading to rapid O_3 destruction within this volcanic plume.

The experimental data reveal the processes behind the observed O_3 depletion, unveiling the intricate interplay between this volcano's emissions and perturbed stratospheric chemistry. Enhanced humidity, radiative cooling, and expanded aerosol surface area in the plume created ideal conditions for rapid chlorine activation on hydrated volcanic aerosols at warm temperatures and subsequent O_3 depletion.

CONCLUSION: Our study sheds light on the complex interactions between a large volcanic eruption and tropical stratospheric O_3, bridging a large gap in our knowledge. The HT eruption's unusual features, notably its high injection altitude of aerosol precursors and large amounts of H_2O vapor, yield invaluable insights into stratospheric chemistry perturbations after this major event.

Beyond its volcanic relevance, our research offers crucial insights into atmospheric chemistry and its implications for climate change. The comprehensive observational evidence that we present substantially advances our understanding of rapid O_3 depletion in certain volcanic plumes. Our study also provides new perspectives on the effects of volcanic eruptions on stratospheric composition and informs future studies and early-response strategies to assess their aftermath. ■



Eruption-triggered rapid O_3 depletion. After the HT eruption, a balloon campaign took place at Réunion Island (left picture). Plume dynamics showcase the volcanic injection of H_2O vapor, sulfur dioxide (SO_2), and HCl, prompting rapid chlorine activation on hydrated volcanic aerosol and O_3 depletion in the stratosphere. The 22 January 2022 O_3 profile (black line) contrasts with Réunion's climatology (red line), displaying a notable decline.

The list of author affiliations is available in the full article online.

*Corresponding author. Email: stephanie.evan@univ-reunion.fr

Cite this article as S. Evan *et al.*, *Science* **382**, eadg2551 (2023). DOI: 10.1126/science.adg2551

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adg2551

RESEARCH ARTICLE

STRATOSPHERIC OZONE

Rapid ozone depletion after humidification of the stratosphere by the Hunga Tonga Eruption

Stephanie Evan[1]*, Jerome Brioude[1], Karen H. Rosenlof[2], Ru-Shan Gao[2], Robert W. Portmann[2], Yunqian Zhu[2,3], Rainer Volkamer[4], Christopher F. Lee[4], Jean-Marc Metzger[5], Kevin Lamy[1], Paul Walter[6], Sergio L. Alvarez[7], James H. Flynn[7], Elizabeth Asher[2,3], Michael Todt[8], Sean M. Davis[2], Troy Thornberry[2], Holger Vömel[9], Frank G. Wienhold[10], Ryan M. Stauffer[11], Luis Millán[12], Michelle L. Santee[12], Lucien Froidevaux[12], William G. Read[12]

The eruption of the Hunga Tonga–Hunga Ha'apai volcano on 15 January 2022 offered a good opportunity to explore the early impacts of tropical volcanic eruptions on stratospheric composition. Balloon-borne observations near Réunion Island revealed the unprecedented amount of water vapor injected by the volcano. The enhanced stratospheric humidity, radiative cooling, and expanded aerosol surface area in the volcanic plume created the ideal conditions for swift ozone depletion of 5% in the tropical stratosphere in just 1 week. The decrease in hydrogen chloride by 0.4 parts per million by volume (ppbv) and the increase in chlorine monoxide by 0.4 ppbv provided compelling evidence for chlorine activation within the volcanic plume. This study enhances our understanding of the effect of this unusual volcanic eruption on stratospheric chemistry and provides insights into possible chemistry changes that may occur in a changing climate.

Large, explosive volcanic eruptions can affect the climate by injecting gases such as sulfur dioxide (SO_2), water vapor (H_2O), and carbon dioxide (CO_2) and halogen compounds such as hydrochloric acid (HCl) and hydrobromic acid (HBr) into the stratosphere. Typically, sulfate aerosols form from sulfur dioxide (SO_2) on a timescale of 1 month (*1*). These aerosols cool the surface by reflecting some incoming solar radiation and warm the stratosphere through the absorption of long-wave radiation. Sulfate aerosol heating can alter stratospheric transport, affecting the distribution of stratospheric species (*1*). The increase in the aerosol burden also enhances aerosol surface area density, increasing the rate of heterogeneous chemical reactions that play a role in stratospheric ozone (O_3) chemistry.

Previous volcanic eruptions had a noticeable impact on stratospheric O_3; examples include El Chichón in 1982 (*2*, *3*) and Mt. Pinatubo in 1991 (*4*, *5*). After the eruption of Mt. Pinatubo, the total O_3 column was reduced by ~6% in northern polar and midlatitude regions (*6*). The O_3 was destroyed by chemical reactions on volcanic aerosols transported to high latitudes, which provided additional reactive surface area for the same heterogeneous chemical reactions that cause polar O_3 loss each year. In the tropics, a 6% decrease in total O_3 was observed in satellite and ozonesonde data (*7*). It was confined between 24 and 28 km in altitude and started 1 month after the eruption, consistent with the time needed for SO_2 to be converted to sulfate aerosols (*8*).

To assess the atmospheric impacts of volcanic eruptions, in situ observations of H_2O vapor, aerosol size and concentration, trace gases (e.g., O_3 and chlorinated species), and meteorological conditions (e.g., temperature and winds) after the eruptions are important (*9*). Major eruptions that cause large perturbations to the aerosol loading and O_3 chemistry of the stratosphere occur sporadically, so the opportunity to make in situ measurements in these plumes is rare.

The Hunga Tonga–Hunga Ha'apai submarine volcano (hereafter HT) in the tropical South Pacific (20.55°S, 175.38°W) was the largest volcanic eruption in the past 30 years. The massive eruption on 15 January 2022 sent material as high as 50 to 55 km, the greatest height ever seen by satellites for a volcanic plume (*10*, *11*). Satellite measurements indicated that the HT volcano injected into the stratosphere a relatively small amount of SO_2, 0.4 Tg (*11*), compared with 10 to 20 Tg from Mt. Pinatubo (*12*, *13*). However, unlike past eruptions, it injected a massive amount of H_2O [~150 Tg, or 10% of the total global mean stratospheric burden (*11*)]. The large amount of injected H_2O was confirmed by radiosonde in situ measurements (*14*).

Less than 1 week after the eruption, a rapid response balloon campaign took place at the Maïdo Observatory (MO) in Réunion Island (21°S, 55°E) (*15*). Because of zonal easterly winds in the stratosphere (20 to 50 km), Réunion Island was ideally located downwind of the plume. From 20 to 24 January 2022, multiple meteorological balloons carrying aerosol, H_2O, SO_2, and O_3 instruments were launched each night to provide key measurements of the volcanic plume's composition (see the materials and methods). A field campaign resulting in volcanic plume measurements in the stratosphere by in situ instruments has never occurred this quickly after an eruption; as a comparison, a rapid response balloon campaign was deployed to Hawaii within 15 days of the 2018 Kilauea eruption (*9*). Herein, we present unprecedented observations corroborating the occurrence of fast chemical depletion of O_3 inside the volcanic plume in the week after the HT eruption. The extremely high H_2O vapor values accelerated the oxidation of SO_2 to sulfate aerosols, with much of the conversion in the densest parts of the plume occurring within a few days after the eruption (*16*), thus facilitating heterogeneous chemical O_3 loss. Additionally, as noted by Anderson *et al.* (*17*), high stratospheric H_2O vapor can change the catalytic chlorine free radical chemistry by shifting total available inorganic chlorine (principally HCl and chlorine nitrate, $ClONO_2$) into the catalytically free-radical form, ClO, potentially increasing O_3 loss. Both rapid aerosol conversion and additional O_3 loss appeared very rapidly for the HT eruption. The high vertical resolution in situ measurements were critical for observing this loss, which is not easily apparent in the satellite O_3 observations, which smear features in the vertical.

Results

Balloon-borne profiles of O_3 and H_2O vapor

Figure 1 shows the balloon-borne profiles of O_3 and H_2O at MO for the period 20 to 24 January 2022. It took 6 days for the volcanic plume to reach Réunion at an ~30-km altitude. Balloon measurements 5 days after the eruption indicate typical stratospheric background values of H_2O (4.5 ppmv), an unperturbed Junge layer backscatter signal, and a peak in O_3 partial pressure of 15 mPa at 25 km. At Réunion, the highest observed H_2O mixing ratio of 356 ppmv at 27 km, 21 Coordinated Universal Time (UTC), was reported on 22 January 2022. Larger values near 1000 ppmv were seen in radiosonde measurements closer to the eruption by Vömel *et al.* (*14*). The observed progression of descending altitude with time in

[1]Laboratoire de l'Atmosphère et des Cyclones (LACy), UMR8105, CNRS, Université de La Réunion, Météo-France, Saint-Denis, France. [2]NOAA Chemical Sciences Laboratory, Boulder, CO, USA. [3]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA. [4]Department of Chemistry and Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA. [5]Observatoire des Sciences de l'Univers de la Réunion, UAR 3365 (CNRS, Université de la Réunion, Météo-France), Saint-Denis, France. [6]St. Edward's University, Austin, TX, USA. [7]University of Houston, Houston, TX, USA. [8]Finnish Meteorological Institute, Helsinki, Finland. [9]National Center for Atmospheric Research, Boulder, CO, USA. [10]Swiss Federal Institute of Technology (ETH), Zurich, Switzerland. [11]Atmospheric Chemistry and Dynamics Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, USA. [12]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA.
*Corresponding author. Email: stephanie.evan@univ-reunion.fr

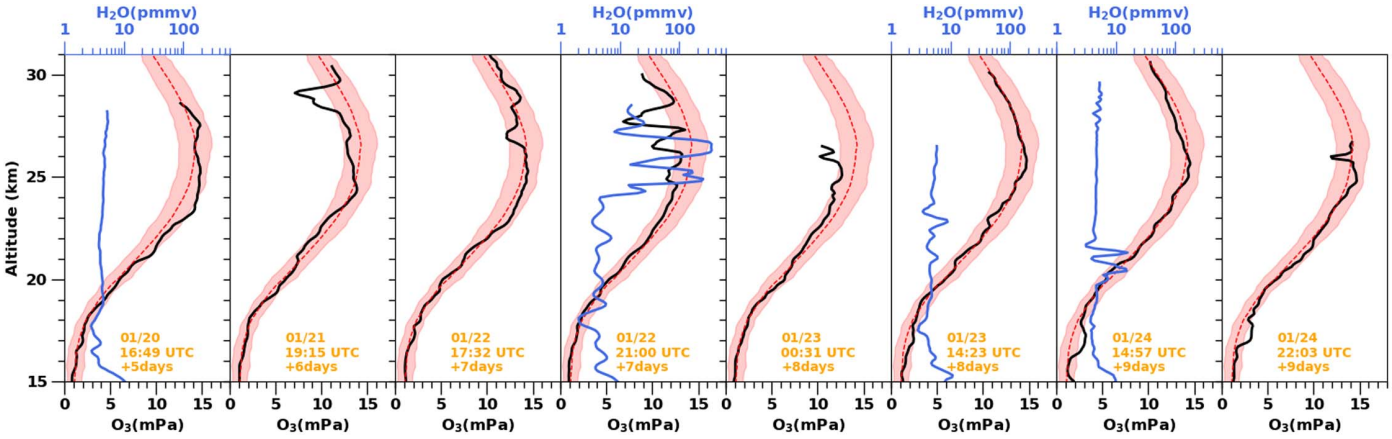


Fig. 1. Soundings of O_3 and H_2O vapor 1 week after the eruption. Profiles of O_3 partial pressure (black) and H_2O vapor mixing ratio (blue) measured at MO on Réunion Island over the period 20 to 24 January 2022. SHADOZ climatological O_3 partial pressure for January for Réunion Island is indicated by the red dashed line. The red shaded area shows values ±1 SD above and below the mean.

the peak of the H_2O plume in Fig. 1 can be explained by the vertical sheer in easterly zonal winds. The easterly zonal winds linearly decreased from 30 m s^{-1} at 30 km to 20 m s^{-1} at 20 km, causing the plume to travel ~18° in longitude per day (*11*, *14*).

The largest O_3 decrease was observed on 21 and 22 January at 25 to 29 km. The decrease amounted to a loss of 10 to 45% relative to O_3 climatology from 1998 to present (see the materials and methods). The presence of SO_2 inside the volcanic plume could affect the performance of electrochemical concentration cell (ECC) ozonesondes by decreasing the sensor signal, resulting in artificially low O_3 measurements. During the experiment, SO_2 instruments (*18*) were launched in tandem with ECC ozonesondes on three flights (see the materials and methods). Overall, the data indicate that SO_2 interference contributes to only a small negative bias of 3 to 4% in the O_3 decrease observed at >25 km in the week after the eruption (see the materials and methods).

Satellite measurements of gas-phase constituents

Satellite measurements of stratospheric gas profiles from the Aura Microwave Limb Sounder (MLS), in operation since 2004 (*19*), were also used to diagnose the possible mechanisms of O_3 decrease after the eruption. Because MLS measures microwave thermal emission from the Earth's limb, its retrievals are largely unaffected by the presence of volcanic aerosols that strongly impair visible, ultraviolet, and infrared measurements. MLS observed the massive H_2O perturbation, as well as anomalous values of several trace gases injected into the stratosphere by HT, such as SO_2 and HCl (*11*). Large H_2O values >100 SDs (σ) above background levels were observed in the week after the eruption at between 20 and 30 km, with a maximum value as high as 350 ppmv at 30 km on 16 January 2022. As shown in Fig. 2, MLS measurements of O_3, HCl, and ClO radicals inside the volcanic plume were identified by selecting data points with H_2O vapor mixing ratios >10 ppmv between 100 and 10 hPa (see the materials and methods). Then, for each data point, the climatological average profile for the month of January from 1998 to 2021 was computed by using MLS profiles in a 5° × 5° box region around the point (see the materials and methods). MLS O_3 mixing ratios inside the volcanic plume decreased at 18 to 21 hPa (~27 to 28 km) in the week after the HT eruption, from an average value of 6.3 ppmv on 16 January 2022 to a minimum value of 4.6 ppmv on 20 January 2022 (near 18°S, 88°E) at 21 hPa, which is 27% below climatological values.

Sensitivity of O_3 decrease to H_2O conditions

We computed average O_3 anomaly and SE profiles inside the plume for the period 16 to 24 January 2022 for two H_2O thresholds, 10 and 100 ppmv, to assess the sensitivity of stratospheric O_3 decrease to H_2O conditions (Fig. 2C). The decrease in MLS O_3 was more pronounced for higher H_2O mixing ratios, and the largest average O_3 drop of 0.4 ppmv (SE of 0.04 ppmv) at 28 km was observed for profiles with H_2O mixing ratios >100 ppmv. Thirty percent of the 190 O_3 profiles selected using the 100 ppmv H_2O mixing ratio threshold showed O_3 dropping below 0.5 ppmv. This change is larger than the typical range of variability established by MLS O_3 measurements for the month of January from 2005 to 2021 period (SD = 0.25 ppmv) or the 2-σ uncertainty of individual MLS O_3 data points (0.1 ppmv) (*20*).

The average O_3 decrease in the ozonesonde profiles was 0.6 ppmv (SE = 0.25 ppmv) near 27 to 29 km, and the strongest decrease was observed during the night of 22 January 2022, when H_2O vapor mixing ratios were the highest. Thus, the large quantity of H_2O vapor injected by the HT volcano appears to have played a key role in the stratospheric O_3 decrease observed by both ozonesondes and MLS.

Discussion

Comparing the HT eruption with that of Mt. Pinatubo

The June 1991 Mt. Pinatubo eruption had a substantial impact on the tropical O_3 column, resulting in a reduction of 13 to 20 Dobson units (DU) within 3 to 6 months (*7*, *8*). This O_3 decrease is comparable to the 18.3 DU loss observed in the 22 January 2022 O_3 profile, but the latter occurred much more rapidly after the HT eruption, indicating the involvement of different mechanisms. The observed low O_3 levels in the tropics after the Mt. Pinatubo eruption were likely caused by increased tropical upwelling due to the heating of volcanic aerosols, facilitating the vertical transport of O_3-depleted air from the troposphere to the stratosphere (*4*). However, the magnitude of O_3 decrease observed after the HT eruption cannot be attributed to lofting, because the radiative cooling effect of the injected H_2O into the stratosphere offset the heating caused by volcanic aerosols (*21*).

Understanding O_3 changes in the tropical stratosphere

A volcanic plume ascending into the stratosphere can incorporate O_3-poor tropospheric air and lead to a localized reduction in column O_3. This effect would have been most prominent in the initial observations of the plume by MLS on 16 January 2022, and over time, the dilution of the volcanic plume would have diminished this apparent loss. However, the minimum O_3 levels were observed at 27 to 28 km altitude 5 days after the eruption, indicating that chemistry played a more prominent role in the observed O_3 changes.

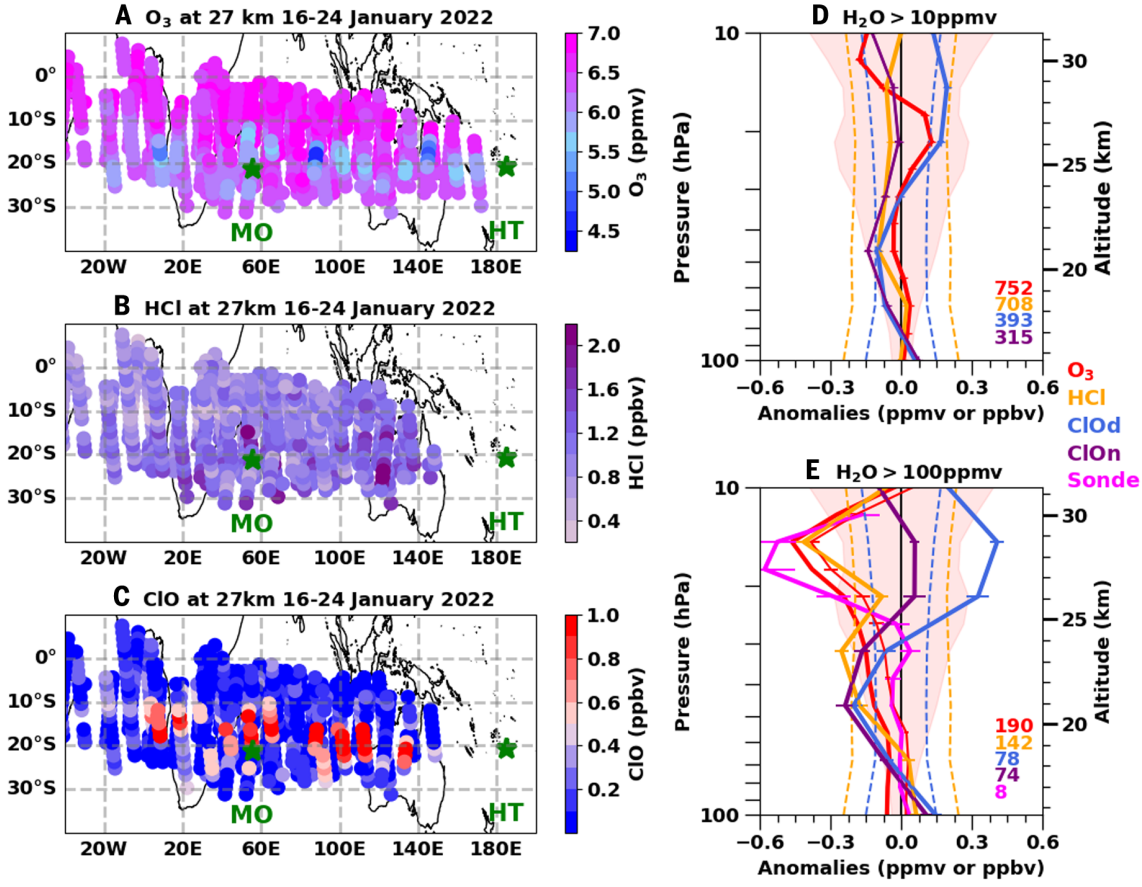


Fig. 2. MLS satellite measurements inside the volcanic plume. (**A** to **C**) MLS maps of O_3 (A), HCl (B), and ClO (C) at 27 km for the period 16 to 24 January 2022. The locations of HT and MO are indicated by green stars. Colored circles correspond to individual satellite profiles. (**D**) MLS anomaly profiles (see the materials and methods) of O_3 (in ppmv; red), HCl (in ppbv; orange), and ClO (in ppbv; blue and purple) inside the volcanic plume averaged over the period 16 to 24 January 2022. Anomalies were computed for MLS measurements inside the plume using a H_2O vapor threshold of 10 ppmv. (**E**) Same as (D) for but H_2O mixing ratios exceeding 100 ppmv. The red shaded area in (D) and (E) indicate ±1 SD of the O_3 climatology. In (E), the average O_3 anomaly profile measured by the balloon sondes in Réunion Island is shown in magenta (ozonesonde data degraded to MLS vertical resolution; see the materials and methods). In (D) and (E), the dashed orange lines correspond to ±1 SD of the HCl climatology, and the dashed blue lines correspond to ±1 SD of the ClO climatology; numbers indicate how many profiles were used to compute the average anomaly. For ClO, the averaged anomaly profiles were computed for daytime (blue) and nighttime (purple) ClO measurements.

The O_3 depletion within the HT volcanic plume exceeded rates observed in previous major eruptions such as El Chichòn and Mt. Pinatubo. Less than 1 week after the HT eruption, O_3 decreased by 0.5 ppmv (i.e., ~0.07 ppmv per day) between 25 and 29 km altitude. The largest seasonal stratospheric O_3 depletion occurs in the polar regions, specifically over the Arctic and Antarctica (*22*). Over the winter months (January to April) in the Arctic, O_3 depletion estimates range from 0.2 to 1.9 ppmv, whereas in Antarctica during the corresponding season, O_3 loss estimates range from 2.3 to 3 ppmv (*23*). These O_3 loss estimates are much larger than the one observed after the HT eruption, although polar O_3 depletion spanned a longer period of 3 to 4 months. During the formation of the Antarctic O_3 hole, stratospheric O_3 is destroyed at a maximum rate of 0.08 ppmv per day (*24*), which is comparable to the rate of 0.07 ppmv per day observed after the HT eruption. Although the comparison involves processes occurring in different regions, it provides valuable context to underscore the substantial and swift O_3 loss observed after the HT eruption.

The stratospheric chemistry behind O_3 depletion

Both gas-phase and heterogeneous chemistry play substantial roles in understanding O_3 depletion in the stratosphere (*22*). Gas-phase chemistry involves a complex suite of reactions between O_3 and various catalytic cycles, notably hydrogen (HOx: hydroxyl radical OH and perhydroxyl radical HO_2), nitrogen (NOx: nitrogen monoxide NO and nitrogen dioxide NO_2), chlorine (ClOx: chlorine Cl and chlorine monoxide ClO), and bromine (BrOx: bromine Br and bromine monoxide BrO) oxides (see the materials and methods). The chemistry involving NOx is closely intertwined with that of halogens (ClOx and BrOx) and of HOx (see the materials and methods).

Reactive halogens, such as chlorine (ClO, Cl) and bromine (BrO, Br) species, have been identified as significant agents in O_3 depletion (*25*). Brominated species, in the presence of elevated Cl, demonstrate higher O_3 destruction efficiency compared with chlorine species due to their longer reaction chains (*26*). The generation of reactive halogen gases involves the transformation of halogen reservoir source gases (HCl, $ClONO_2$, HBr, and $BrONO_2$). Photolysis plays a key role in dissociating bromine reservoir species (HBr, $BrONO_2$), whereas chlorine reservoir species (HCl, $ClONO_2$) undergo relatively slow photolysis in the gas phase. However, more rapid transformations occur through heterogeneous reactions on solid or liquid particle surfaces (see the materials and methods).

Heterogeneous chlorine chemistry has been well established in polar regions, where polar stratospheric clouds form at temperatures below ~195 K. These clouds, along with cold sulfate particles, play a crucial role in the conversion of inorganic chlorine species ($ClONO_2$ and HCl) into chemically active forms of chlorine, namely ClO and Cl (*22*), which subsequently initiate catalytic O_3 loss in the polar regions. Activation of $ClONO_2$ and HCl through heterogeneous reactions becomes more efficient as temperature decreases or as increased H_2O dilutes the H_2SO_4 concentration (*27*).

Some volcanic eruptions have been observed to emit halogens such as HCl and HBr into the stratosphere, along with SO_2 (*28*). Halogens, particularly HCl, are highly soluble and can be scavenged by H_2O, ice hydrometeors, and ash present in the volcanic plume (*29*). Theoretical predictions and advanced plume models indicate that 10 to 20%, of the HCl emitted at the vent during large, explosive volcanic eruptions could directly reach the stratosphere (*28*). Thus, observational examples of significant HCl and HBr emissions from volcanic events that directly lead to stratospheric O_3 depletion are limited (*30*). Depletion of stratospheric O_3 observed after volcanic eruptions has been mainly attributed to heterogeneous reactions occurring on volcanic sulfate aerosols, based on modeling studies [e.g., (*2*, *31*)]. Furthermore, the increased presence of stratospheric sulfate aerosols resulting from volcanic eruptions can also lead to a reduction in the abundance of NOx (*32*), making the ClOx and HOx cycles more effective at destroying O_3 (see the materials and methods). If less NOx is available, then the catalytic O_3 depletion through NOx could slow down. However, this effect has primarily been observed at higher altitudes (>30 km) after the Mt. Pinatubo eruption (*33*).

Although the impact of volcanic eruptions on O_3 depletion in midlatitude regions is relatively well documented [e.g., (*4*, *34*)], there have been fewer studies that specifically focus on the effects of volcanic eruptions and heterogeneous chemistry on the distribution of stratospheric O_3 in tropical regions, because most volcanic injections do not reach altitudes high enough to substantially influence stratospheric O_3 chemistry in the tropics. However, the case of the HT volcanic eruption is different because the injection altitude was exceptionally high.

Role of H_2O vapor and temperature

Most of the evidence for the occurrence of heterogeneous chemistry in the midlatitude and tropical stratosphere was provided for the tropopause region, where low temperatures are encountered. Modeling studies have indicated the possibility of heterogeneous chemistry occurring on cirrus cloud particles near the midlatitude tropopause (*35*, *36*) or heterogeneous chlorine reactions on liquid sulfate aerosols near tropical monsoon regions (*37*).

Modeling studies have also indicated that enhanced H_2O in the stratosphere could lead to chemical O_3 depletion through heterogeneous chlorine activation and subsequent O_3 destruction (*38*, *39*). Low temperatures (<203 K) and elevated H_2O mixing ratios (>20 ppmv) must be maintained for heterogeneous chlorine activation on sulfate aerosols to occur (*39*).

The largest O_3 loss observed above Réunion Island on 22 January 2022 coincides with H_2O mixing ratios exceeding 100 ppmv at ~25 and 27 km (Fig. 3). Above 25 km, temperature values were 2 to 4 K below the climatological mean temperature profile for January computed using 1998–2021 SHADOZ data for Réunion Island (see the materials and methods). Radiosoundings in the tropics have indicated an average temperature decrease of 2 K caused by the large increase of H_2O vapor after the HT eruption (*14*). Figure 3 shows that O_3 decreases also coincide with the presence of aerosols, as indicated by peaks in the COBALD backscatter measurements. Measurements from an optical particle counter instrument flown that night provide additional information on the aerosol size and concentration, from which aerosol surface area density can be estimated (see the materials and methods). In Fig. 3, the aerosol measurements indicate a peak ambient surface area density of 286.9 $\mu m^2\ cm^{-3}$ at 26.4 km. This is 4 to 6 times higher than what was reported 1 month after the Mt. Pinatubo eruption (*40*) and ~600 times higher than the

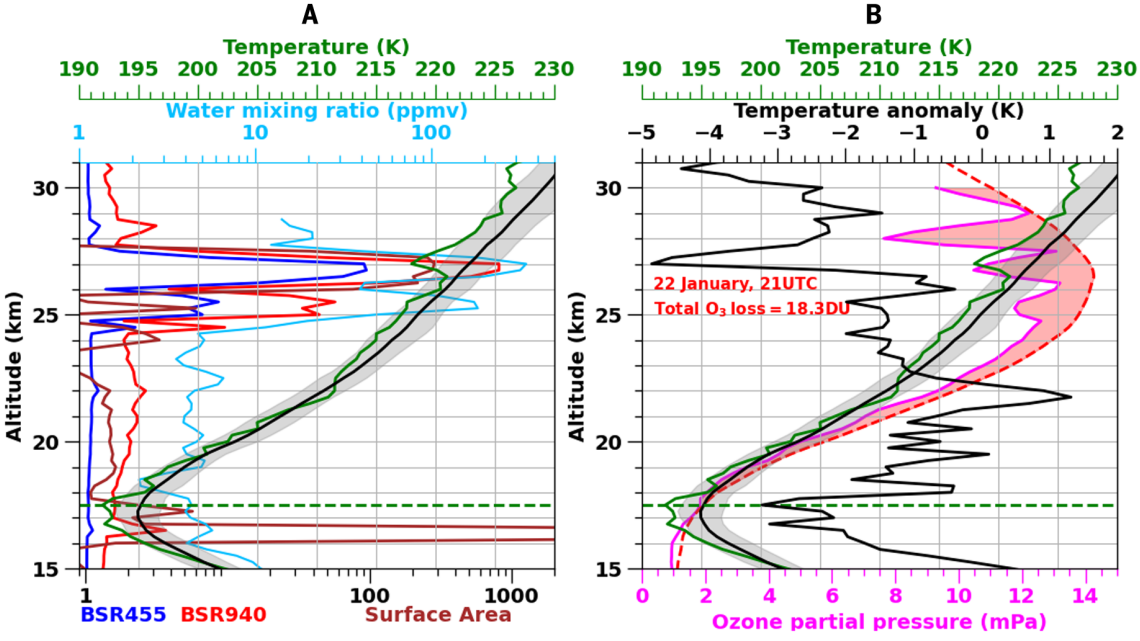


Fig. 3. Soundings at MO on 22 January 2022. (**A** and **B**) Profiles of H_2O vapor mixing ratio (blue), aerosol backscatter ratios at 940 nm (red) and 455 nm (blue), aerosol ambient surface area density (brown, in $\mu m^2\ cm^{-3}$) (A), and O_3 partial pressure (magenta) and temperature (green) (B) measured at MO on 22 January 2022. The horizontal green dashed line shows the location of the tropopause, and the red dashed line shows the O_3 climatology for Réunion Island. The red shaded area illustrates the inferred O_3 loss estimated for that night. The climatological temperature profile for January 1998 to 2021 for Réunion Island is shown in black in (A) and (B); the temperature anomaly from the climatology is shown in (B). The gray shaded area illustrates ±1 SD of the temperature climatology. The average temperature anomaly was −2.2 K between 25 and 30 km.

background stratospheric value of 0.5 μm^2 cm^{-3}. The large humidification of the stratosphere by the HT eruption resulted in rapid aerosol formation in the volcanic plume and in the corresponding large aerosol surface area (*16*).

MLS measurements at 27 km inside the volcanic plume indicate that O_3 decreases (average O_3 anomaly of −0.4 ppmv with decreases as large as −1.5 ppmv) correlated with higher H_2O vapor mixing ratios (mean H_2O of 102 ppmv) and temperatures on average ~2 K below the climatological average (see the materials and methods and fig. S1). Figure 2 shows that a decrease in HCl of 0.4 per million by volume (ppbv) at 28 km corresponds to an increase in ClO of 0.4 ppbv. Low HCl and high ClO reflect the conversion of the halogen reservoir compounds, such as HCl and $ClONO_2$, to the reactive form of chlorine, ClO. Above 25 km, where most of the O_3 loss is observed, it is usually far too warm (>220 K) for heterogeneous chlorine activation to occur (see the materials and methods). However, the large stratospheric humidification, subsequent radiative cooling, and the added surface area after the HT eruption likely accelerated heterogeneous chlorine activation on sulfate aerosols and led to O_3 decreases. To illustrate the impacts of H_2O vapor increase and subsequent radiative cooling on chlorine heterogeneous chemistry, we considered the heterogeneous reactions of $ClONO_2$ + H_2O, $ClONO_2$ + HCl, and HOCl + HCl on sulfate aerosols. The rates of heterogeneous reactions on sulfate aerosols are highly influenced by temperature and H_2O vapor partial pressure, because these factors determine the composition of the sulfate aerosols and the solubility of HCl, $ClONO_2$, and HOCl (*41*).

When H_2O vapor mixing ratios exceed 100 ppmv, the probabilities of the heterogeneous reactions $ClONO_2$ + HCl and HOCl + HCl increase by four orders of magnitude under warm conditions (see the materials and methods and fig. S2). The radiative cooling induced by the large amount of H_2O vapor (−4 K; Fig. 3) can further increase these probabilities by another factor of 10 (see the materials and methods and fig. S2). An increase in H_2O vapor and a corresponding decrease in temperature results in more dilute sulfate aerosols and higher HCl solubility; therefore, the reaction probability for the conversion of HCl to active ClOx (Cl, ClO) increases markedly, and catalytic O_3 loss becomes possible under relatively warm conditions.

Evidence of chlorine activation on hydrated volcanic aerosols

Although the HCl injection into the stratosphere by the HT eruption was comparable to that from previous moderate eruptions observed during the MLS record (*11*), it reached an altitude near 24 km (31.6 hPa) on 16 to 17 January 2022, which is well above levels previously observed by MLS (*11*). In Fig. 2, two main negative HCl anomalies are seen at 24 and 29 km for measurements with large H_2O vapor mixing ratios (>100 ppmv) inside the volcanic plume. The first HCl anomaly value of −0.25 ppbv at 24 km may have been due to scavenging by ice or ash particles (*29*), which were injected up to ~23 km but rapidly washed out within the first day after the eruption (*42*). The second, larger HCl anomaly of −0.4 ppbv at 29 km may indicate HCl heterogeneous activation on sulfate aerosols. Indeed, this negative HCl anomaly coincides with a positive peak of 0.41 ppbv in ClO radicals that is much more pronounced for daytime measurements (Fig. 2). This is consistent with photolysis of Cl_2 and the subsequent reaction of Cl with O_3 to form ClO.

After the HT eruption, elevated H_2O vapor levels reaching at least 70 times the background levels were observed. These increased H_2O concentrations can play a crucial role in enhancing gas-phase chemistry involving HOx radicals, particularly OH and HO_2. The higher H_2O concentrations may boost the HOx cycle, specifically affecting reactions such as OH + O_3 and HO_2 + O_3, which in turn can lead to O_3 depletion. Additionally, the rise in H_2O concentrations strengthens the interaction between the HOx and ClOx cycles (through HO_2 + ClO → HOCl + O_2), with potential implications for O_3 depletion processes (*39*).

A decrease in the observed NO_2 radical abundance indicates slowing of the NO_x cycle, presumably due to N_2O_5 hydrolysis (see the materials and methods and fig. S3). At sunset, the NO_2 concentration is higher than in the morning and better approximates baseline conditions, suggesting that NO_x may at least partially be sustained in the daytime plume. This reduction in NOx can reinforce the ClOx and HOx cycles, contributing to O_3 depletion.

BrO and OClO have been detected in volcanic plumes in the upper troposphere to lower stratosphere region (*43*, *44*), indicating the activation of bromine and chlorine species (because the ClO + BrO reaction is the primary known source of OClO). Li *et al.* (*45*) reported SO_2 and BrO vertical column densities from satellite measurements in the HT volcanic plume for the period 15 to 19 January 2022. Most of the injected BrO spread southeast of the volcano on 16 January 2022, whereas a smaller portion continued to propagate eastward in the subsequent days. Ground-based differential optical absorption spectroscopy (DOAS) observations cannot confirm the presence of elevated BrO in the volcanic plume beyond 19 January 2022 (see the materials and methods), when it was last reported near 130°E in northeastern Australia (*45*). The potential presence of BrO, along with the MLS detection of ClO in the volcanic plume, suggests the possibility of O_3 depletion in the volcanic cloud through the ClO–BrO catalytic cycle (*46*) (see the materials and methods).

Insights from modeling

A recent study by Zhu *et al.* (*47*) used the Whole Atmosphere Community Climate Model (WACCM) to investigate the role of chemistry and transport in O_3 depletion after the HT eruption. The authors simulated the conditions observed by MLS after the eruption by injecting H_2O, ClO (or HCl), and SO_2 into the model. One of their simulations focused on the impact of injecting low O_3, which explained the initial low values observed. However, it was found that these low O_3 levels did not persist over time, suggesting that the injection of tropospheric air poor in O_3 was not a substantial contributing factor. It was also noted that inside the volcanic plume, the concentrations of OH and HO_2 increased by a factor of 10 compared with background levels, with a corresponding enhancement in reaction rates by a factor of 10, specifically for the O_3 loss reactions HO_2 + O_3 and OH + O_3 (see the materials and methods).

Zhu *et al.* also confirmed a large enhancement, by several orders of magnitude, of the heterogeneous reaction rates for $ClONO_2$ + HCl, $ClONO_2$ + H_2O, and HOCl + HCl (see the materials and methods) (*47*). They found that HOCl + HCl was a major cause of chlorine activation. Additionally, they calculated a 20-fold increase in the reaction rate of ClO + HO_2 → HOCl + O_2 (see the materials and methods). The authors also concluded that the massive H_2O injection during the HT eruption led to significant changes in ClOx–HOx interactions and heterogeneous reaction rates, resulting in diverse chemical pathways for chlorine activation and O_3 depletion. HOCl played a crucial role in both in-plume chlorine balance and heterogeneous processes. Both gas-phase chemistry and heterogeneous chemistry are deemed significant factors contributing to O_3 depletion (*47*).

These strengthened ClOx and ClO/HO_2 cycles led to an increase in the O_3 loss reaction Cl + O_3 by a factor of 2, which is sufficient to account for the O_3 loss observed by MLS, even if the contribution from the bromine catalytic cycle is not considered.

Potential mechanisms and impacts on O_3 depletion

The large amount of seawater vaporized during the HT eruption contained sea salt that may have been transported to the stratosphere, providing a source of reactive chlorine, bromine, and iodine species for O_3 destruction. However, it is unclear whether all components of sea salt aerosols would survive wet scavenging in the volcanic plume and be efficiently transported to the stratosphere. Inorganic chlorine (HCl, Cl^-) and bromine (HBr, Br^-) are easily removed by

wet scavenging, whereas inorganic iodine, the dominant form of stratospheric iodine injections, is removed less efficiently (*48*). Even a small injection of iodine into the stratosphere could have accelerated O_3 loss in the volcanic plume because of its much larger O_3-depleting efficiency (*48*–*50*).

Anderson *et al.* (*38*) hypothesized that the excess midlatitude stratospheric H_2O associated with convection changes caused by global warming could destroy lower stratospheric O_3. This HT event appears to be an extreme example of that mechanism. The quantity of H_2O from HT is significantly larger than would be expected from overshooting monsoon convection but provides observational evidence that this process could occur (see the materials and methods). The rapid chemical O_3 loss inside the HT volcanic plume was primarily triggered by the synergistic effects of large humidification, radiative cooling, and added aerosol surface area. All of these effects acted in concert to accelerate HCl heterogeneous activation on sulfate aerosols at warmer temperatures than have been recorded previously. The fact that stratospheric O_3 loss was so rapid after the HT eruption calls for additional experimental studies of heterogeneous chemistry on hydrated aerosols (especially halogen chemistry) to better assess the kinetics of these reactions. It also identifies the need to develop atmospheric modeling tools at finer scales to better understand stratospheric O_3 chemistry in the tropics after extreme events. The increased stratospheric H_2O may linger for 4 to 5 years, potentially altering O_3 chemistry. As noted in Solomon *et al.* (*51*) in regard to the Australian New Year bush fires in 2019–2020, which injected 0.9 Tg of smoke into the stratosphere, there were chemical shifts associated with heterogeneous reactions that altered nitrogen, chlorine and reactive hydrogen species with the potential to cause midlatitude O_3 loss. For the next few years, continued monitoring of O_3, H_2O vapor, aerosols, and halogens is needed to assess the extended impacts of this unprecedented HT eruption.

Methods summary

In situ balloon-borne measurements

Balloon-borne measurements at the MO in Réunion Island in January 2022 included instruments such as the Cryogenic Frost-Point Hygrometer (CFH), the Compact Optical Backscatter and Aerosol Detector (COBALD), the Electrochemical Concentration Cell (ECC) Ozonesonde, SO_2 sonde, and the Portable Optical Particle Counter (POPS). The CFH instrument measures water vapor with an uncertainty of 5 to 10% or less in the stratosphere. COBALD characterizes aerosol particles and clouds based on the backscatter ratio at two wavelengths. ECC Ozonesonde provides O_3 profile data with an uncertainty of 10%. The SO_2 sonde provides SO_2 profile data with an uncertainty of 20%. The Intermet iMet-4-RSB Meteorological Radiosonde supports data telemetering to the ground. POPS measures aerosol size distribution with an uncertainty of 5%. These measurements have a high vertical resolution of 5 m, and all variables were binned in altitude intervals of 100 m to reduce measurement noise.

MLS observations and comparison to high-resolution ozonesonde measurements

The Aura MLS satellite instrument observes thermal emissions from the atmosphere over a wide latitude range. MLS measurements within the volcanic plume were identified by selecting data points with water vapor mixing ratios >10 ppmv and between 100 and 10 hPa. Quality screening criteria were applied to the MLS data to ensure accuracy. To compare high-resolution balloon-borne ozonesonde measurements with MLS data, the sonde data were smoothed to match MLS resolution, and MLS vertical averaging kernels were applied using a priori climatological profiles. Overall, this approach enables the comparison of high-resolution balloon-borne measurements with MLS satellite data while accounting for resolution differences and other factors.

Ozonesonde and SO_2 interference

The presence of SO_2 in the atmosphere can interfere with ECC ozonesonde measurements. ECC ozonesondes measure an electrical current that is produced when O_3 enters the sensor and reacts with KI to produce I within the sonde. By measuring the electron flow and the rate at which O_3 enters the sensor per unit time, the O_3 concentration can be calculated. However, when SO_2 is present, it can convert I^- back into I, leading to lower measured O_3 levels. During the rapid response experiment at MO, SO_2 instruments were launched alongside ECC ozonesondes on some payloads. The 21 January 2022 flight reported that SO_2 interference accounted for ~3 to 4% of the observed O_3 decrease in the ozonesonde profile between 28 and 30 km in altitude.

Zenith-Sky DOAS measurements of stratospheric NO_2

Measurements of stratospheric NO_2 from the University of Colorado DOAS instrument during the plume period (21 January to 1 February 2022) were compared with the clear-sky interannual NO_2 variability for 3 years (January 2020 to January 2022) above MO to assess whether the NOx cycle of stratospheric O_3 destruction was significantly affected by dinitrogen pentoxide (N_2O_5) hydrolysis in the humidified plume. The instrument presents qualitative evidence for a reduction in stratospheric NO_2 during sunrise and sunset, consistent with N_2O_5 hydrolysis, in both geometries overlapping with the plume. However, no evidence for enhanced BrO due to volcanic plume injection was detectable in the DOAS data.

REFERENCES AND NOTES

1. J. H. Seinfeld, S. N. Pandis, *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change* (Wiley, ed. 2, 2006).
2. D. J. Hofmann, S. Solomon, Ozone destruction through heterogeneous chemistry following the eruption of El Chichón. *J. Geophys. Res.* **94**, 5029–5041 (1989). doi: 10.1029/JD094iD04p05029
3. A. Adriani, G. Fiocco, G. P. Gobbi, F. Congeduti, Correlated behavior of the aerosol and ozone contents of the stratosphere after the El Chichon eruption. *J. Geophys. Res.* **92**, 8365–8372 (1987). doi: 10.1029/JD092iD07p08365
4. G. Brasseur, C. Granier, Mount pinatubo aerosols, chlorofluorocarbons, and ozone depletion. *Science* **257**, 1239–1242 (1992). doi: 10.1126/science.257.5074.1239; pmid: 17742756
5. W. J. Randel, F. Wu, J. M. Russellllll, J. W. Waters, L. Froidevaux, Ozone and temperature changes in the stratosphere following the eruption of Mount Pinatubo. *J. Geophys. Res.* **100**, 16753–16764 (1995). doi: 10.1029/95JD01001
6. J. K. Angell, Estimated impact of Agung, El Chichon and Pinatubo volcanic eruptions on global and regional total ozone after adjustment for the QBO. *Geophys. Res. Lett.* **24**, 647–650 (1997). doi: 10.1029/97GL00544
7. M. R. Schoeberl, P. K. Bhartia, E. Hilsenrath, O. Torres, Tropical ozone loss following the eruption of Mt. Pinatubo. *Geophys. Res. Lett.* **20**, 29–32 (1993). doi: 10.1029/92GL02637
8. W. B. Grant *et al.*, Observations of reduced ozone concentrations in the tropical stratosphere after the eruption of Mt. Pinatubo. *Geophys. Res. Lett.* **19**, 1109–1112 (1992). doi: 10.1029/92GL01153
9. J.-P. Vernier *et al.*, VolKilau: Volcano Rapid Response Balloon Campaign during the 2018 Kilauea Eruption. *Bull. Am. Meteorol. Soc.* **101**, E1602–E1618 (2020). doi: 10.1175/BAMS-D-19-0011.1
10. S. A. Carn, N. A. Krotkov, B. L. Fisher, C. Li, Out of the blue: Volcanic SO2 emissions during the 2021–2022 eruptions of Hunga Tonga—Hunga Ha'apai (Tonga). *Front. Earth Sci. (Lausanne)* **10**, 976962 (2022). doi: 10.3389/feart.2022.976962
11. L. Millán *et al.*, The Hunga Tonga-Hunga Ha'apai Hydration of the Stratosphere. *Geophys. Res. Lett.* **49**, GL099381 (2022). doi: 10.1029/2022GL099381; pmid: 35865735
12. G. J. S. Bluth, S. D. Doiron, C. C. Schnetzler, A. J. Krueger, L. S. Walter, Global tracking of the SO $_2$ clouds from the June, 1991 Mount Pinatubo eruptions. *Geophys. Res. Lett.* **19**, 151–154 (1992). doi: 10.1029/91GL02792
13. W. G. Read, L. Froidevaux, J. W. Waters, Microwave limb sounder measurement of stratospheric SO_2 from the Mt. Pinatubo Volcano. *Geophys. Res. Lett.* **20**, 1299–1302 (1993). doi: 10.1029/93GL00831
14. H. Vömel, S. Evan, M. Tully, Water vapor injection into the stratosphere by Hunga Tonga-Hunga Ha'apai. *Science* **377**, 1444–1447 (2022). doi: 10.1126/science.abq2299; pmid: 36137033
15. J.-L. Baray *et al.*, Maïdo Observatory: A new high-altitude station facility at Reunion Island (21° S, 55° E) for long-term atmospheric remote sensing and in situ measurements. *Atmos. Meas. Tech.* **6**, 2865–2877 (2013). doi: 10.5194/amt-6-2865-2013
16. Y. Zhu *et al.*, Perturbations in stratospheric aerosol evolution due to the water-rich plume of the 2022 Hunga-Tonga eruption. *Commun. Earth Environ.* **3**, 248 (2022). doi: 10.1038/s43247-022-00580-w
17. J. G. Anderson, D. M. Wilmouth, J. B. Smith, D. S. Sayres, UV dosage levels in summer: Increased risk of ozone loss from convectively injected water vapor. *Science* **337**, 835–839 (2012). doi: 10.1126/science.1222978; pmid: 22837384
18. S. Yoon *et al.*, Development and testing of a novel sulfur dioxide sonde. *Atmos. Meas. Tech.* **15**, 4373–4384 (2022). doi: 10.5194/amt-15-4373-2022
19. J. W. Waters *et al.*, The Earth Observing System Microwave Limb Sounder (EOS MLS) on the Aura satellite. *IEEE Trans. Geosci. Remote Sens.* **44**, 1075–1092 (2006). doi: 10.1109/TGRS.2006.873771
20. N. J. Livesey *et al.*, "EOS MLS version 4.2x level 2 data quality and description document (revision E)" (Technical Report, Jet Propulsion Laboratory, NASA, 2020); https://mls.jpl.nasa.gov/data/v4-2_data_quality_document.pdf

21. P. Sellitto *et al*., The unexpected radiative impact of the Hunga Tonga eruption of 15th January 2022. *Commun. Earth Environ.* **3**, 288 (2022). doi: 10.1038/s43247-022-00618-z
22. S. Solomon, Stratospheric ozone depletion: A review of concepts and history. *Rev. Geophys.* **37**, 275–316 (1999). doi: 10.1029/1999RG900008
23. N. J. Livesey, M. L. Santee, G. L. Manney, A Match-based approach to the estimation of polar stratospheric ozone loss using Aura Microwave Limb Sounder observations. *Atmos. Chem. Phys.* **15**, 9945–9963 (2015). doi: 10.5194/acp-15-9945-2015
24. B. J. Johnson *et al*., South Pole Station ozonesondes: Variability and trends in the springtime Antarctic ozone hole 1986–2021. *Atmos. Chem. Phys.* **23**, 3133–3146 (2023). doi: 10.5194/acp-23-3133-2023
25. D. J. Lary, Catalytic destruction of stratospheric ozone. *J. Geophys. Res.* **102**, 21515–21526 (1997). doi: 10.1029/97JD00912
26. R. J. Salawitch, Sensitivity of ozone to bromine in the lower stratosphere. *Geophys. Res. Lett.* **32**, L05811 (2005). doi: 10.1029/2004GL021504
27. M. A. Tolbert, M. J. Rossi, D. M. Golden, Heterogeneous interactions of chlorine nitrate, hydrogen chloride, and nitric acid with sulfuric acid surfaces at stratospheric temperatures. *Geophys. Res. Lett.* **15**, 847–850 (1988). doi: 10.1029/GL015i008p00847
28. C. Textor, Injection of gases into the stratosphere by explosive volcanic eruptions. *J. Geophys. Res.* **108**, 4606 (2003). doi: 10.1029/2002JD002987
29. A. Tabazadeh, R. P. Turco, Stratospheric chlorine injection by volcanic eruptions: HCl scavenging and implications for ozone. *Science* **260**, 1082–1086 (1993). doi: 10.1126/science.260.5111.1082; pmid: 17806335
30. J. Staunton-Sykes *et al*., Co-emission of volcanic sulfur and halogens amplifies volcanic effective radiative forcing. *Atmos. Chem. Phys.* **21**, 9009–9029 (2021). doi: 10.5194/acp-21-9009-2021
31. C. Granier, G. Brasseur, Impact of heterogeneous chemistry on model predictions of ozone changes. *J. Geophys. Res.* **97**, 18015–18033 (1992). doi: 10.1029/92JD02021
32. M. Prather, Catastrophic loss of stratospheric ozone in dense volcanic clouds. *J. Geophys. Res.* **97**, 10187–10191 (1992). doi: 10.1029/92JD00845
33. M. Kilian, S. Brinkop, P. Jöckel, Impact of the eruption of Mt Pinatubo on the chemical composition of the stratosphere. *Atmos. Chem. Phys.* **20**, 11697–11715 (2020). doi: 10.5194/acp-20-11697-2020
34. V. Aquila, L. D. Oman, R. S. Stolarski, P. R. Colarco, P. A. Newman, Dispersion of the volcanic sulfate cloud from a Mount Pinatubo–like eruption. *J. Geophys. Res.* **117**, D06216 (2012). doi: 10.1029/2011JD016968
35. S. Borrmann, S. Solomon, L. Avallone, D. Toohey, D. Baumgardner, On the occurrence of ClO in cirrus clouds and volcanic aerosol in the tropopause region. *Geophys. Res. Lett.* **24**, 2011–2014 (1997). doi: 10.1029/97GL02053
36. S. Solomon *et al*., Heterogeneous chlorine chemistry in the tropopause region. *J. Geophys. Res.* **102**, 21411–21429 (1997). doi: 10.1029/97JD01525
37. S. Solomon *et al*., Monsoon circulations and tropical heterogeneous chlorine chemistry in the stratosphere. *Geophys. Res. Lett.* **43**, (2016). doi: 10.1002/2016GL071778
38. J. G. Anderson *et al*., Stratospheric ozone over the United States in summer linked to observations of convection and temperature via chlorine and bromine catalysis. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E4905–E4913 (2017). doi: 10.1073/pnas.1619318114; pmid: 28584119
39. S. Robrecht *et al*., Mechanism of ozone loss under enhanced water vapour conditions in the mid-latitude lower stratosphere in summer. *Atmos. Chem. Phys.* **19**, 5805–5833 (2019). doi: 10.5194/acp-19-5805-2019
40. T. Deshler, D. J. Hofmann, B. J. Johnson, W. R. Rozier, Balloonborne measurements of the Pinatubo aerosol size distribution and volatility at Laramie, Wyoming during the summer of 1991. *Geophys. Res. Lett.* **19**, 199–202 (1992). doi: 10.1029/91GL02787
41. D. R. Hanson, A. R. Ravishankara, S. Solomon, Heterogeneous reactions in sulfuric acid aerosols: A framework for model calculations. *J. Geophys. Res.* **99**, 3615–3629 (1994). doi: 10.1029/93JD02932
42. B. Legras *et al*., The evolution and dynamics of the Hunga Tonga–Hunga Ha'apai sulfate aerosol plume in the stratosphere. *Atmos. Chem. Phys.* **22**, 14957–14970 (2022). doi: 10.5194/acp-22-14957-2022
43. N. Theys *et al*., Global observations of tropospheric BrO columns using GOME-2 satellite data. *Atmos. Chem. Phys.* **11**, 1791–1811 (2011). doi: 10.5194/acp-11-1791-2011
44. C. Hörmann *et al*., Systematic investigation of bromine monoxide in volcanic plumes from space by using the GOME-2 instrument. *Atmos. Chem. Phys.* **13**, 4749–4781 (2013). doi: 10.5194/acp-13-4749-2013
45. Q. Li *et al*., Diffusion height and order of sulfur dioxide and bromine monoxide plumes from the Hunga Tonga–Hunga Ha'apai volcanic eruption. *Remote Sens. (Basel)* **15**, 1534 (2023). doi: 10.3390/rs15061534
46. M. B. McElroy, R. J. Salawitch, S. C. Wofsy, J. A. Logan, Reductions of Antarctic ozone due to synergistic interactions of chlorine and bromine. *Nature* **321**, 759–762 (1986). doi: 10.1038/321759a0
47. Y. Zhu *et al*., "Stratospheric ozone depletion inside the volcanic plume shortly after the 2022 Hunga Tonga eruption" (EGUsphere, 2023); https://doi.org/10.5194/egusphere-2023-1334.
48. T. K. Koenig *et al*., Quantitative detection of iodine in the stratosphere. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 1860–1866 (2020). doi: 10.1073/pnas.1916828117; pmid: 31932452
49. World Meteorological Organization, "Scientific assessment of ozone depletion: 2018" (WMO, 2018), Global Ozone Research and Monitoring Project, Report No. 58; https://csl.noaa.gov/assessments/ozone/2018/.
50. L. J. Carpenter *et al*., "Update on ozone-depleting substances (ODSs) and other gases of interest to the Montreal Protocol," in *Scientific Assessment of Ozone Depletion: 2014* (World Meteorological Organization, 2014), Global Ozone Research and Monitoring Project Report No. 55; https://www.nist.gov/publications/update-ozone-depleting-substances-odss-and-other-gases-interest-montreal-protocol.
51. S. Solomon *et al*., On the stratospheric chemistry of midlatitude wildfire smoke. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2117325119 (2022). doi: 10.1073/pnas.2117325119; pmid: 35238658
52. CFH ozonesonde data for: S. Evan *et al*., Rapid ozone depletion after humidification of the stratosphere by the Hunga Tonga Eruption, GeoNetwork (2023); https://doi.org/10.26171/p4zy-fy98.
53. ECC ozonesonde data for: S. Evan *et al*., Rapid ozone depletion after humidification of the stratosphere by the Hunga Tonga Eruption, GeoNetwork (2023); https://doi.org/10.26171/mvee-5v84.
54. Raw MLS H_2O data for: S. Evan *et al*., Rapid ozone depletion after humidification of the stratosphere by the Hunga Tonga Eruption, GES DISC (2023); https://disc.gsfc.nasa.gov/datasets/ML2H2O_004/summary?keywords=aura.
55. Raw MLS O_3 data for: S. Evan *et al*., Rapid ozone depletion after humidification of the stratosphere by the Hunga Tonga Eruption, GES DISC (2023); https://disc.gsfc.nasa.gov/datasets/ML2O3_004/summary?keywords=aura.
56. Raw MLS HCl data for: S. Evan *et al*., Rapid ozone depletion after humidification of the stratosphere by the Hunga Tonga Eruption, GES DISC (2023); https://disc.gsfc.nasa.gov/datasets/ML2HCL_004/summary?keywords=aura.
57. Raw MLS ClO data for: S. Evan *et al*., Rapid ozone depletion after humidification of the stratosphere by the Hunga Tonga Eruption, GES DISC (2023); https://disc.gsfc.nasa.gov/datasets/ML2CLO_004/summary?keywords=aura.

ACKNOWLEDGMENTS

We thank G. Morris, D. Murphy, M. Abou-Ghane, and J. Burkholder for discussions; A. Wong, T. Koenig, C. Thompson, and SilverLining for support; and the European Communities, the Région Réunion, CNRS, and Université de la Réunion for support and contributions in the construction phase of the research infrastructure Observatoire de Physique de l'Atmosphère de La Réunion, including Maïdo Observatory (OPAR). OPAR is presently funded by CNRS (INSU), Météo France, and Université de La Réunion and managed by OSU-R (Observatoire des Sciences de l'Univers de La Réunion, UAR 3365). **Funding:** This work was supported by the Agence Nationale de la Recherche (ANR CONCIRTO award ANR-17-CE01-0005-01), the National Science Foundation (NSF awards AGS-1620530 and AGS-2027252), and the National Aeronautics and Space Administration (NASA contract 80NM0018D0004). **Author contributions:** Conceptualization: S.E.; Data curation: S.E., J.B., R.V., C.F.L., J.M.M., K.L., P.W., S.L.A., J.H.F., E.A., L.A., M.T., H.V., F.G.W., L.M., M.L.S., L.F., W.G.R.; Methodology: S.E., J.B.; Writing – original draft: all authors. Writing – review & editing: all authors. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The processed CFH (*52*) and ECC (*53*) ozonesonde data used in this study are available at GeoNetwork. Raw MLS H_2O (*54*), O_3 (*55*), HCl (*56*), and ClO (*57*) data may be found at the NASA Goddard Earth Sciences Data and Information Services Center (GES DISC) repository. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adg2551
Materials and Methods
Figs. S1 to S3
Table S1
References (*58*–*86*)

Submitted 12 December 2022; accepted 5 September 2023
10.1126/science.adg2551

RESEARCH ARTICLE SUMMARY

NEUROSCIENCE

Associative and predictive hippocampal codes support memory-guided behaviors

Can Liu†, Ralitsa Todorova†, Wenbo Tang, Azahara Oliva, Antonio Fernandez-Ruiz*

INTRODUCTION: The brain generates models of the environment that are used to guide flexible behaviors. This process requires learning the states of the world (such as specific locations) as well as the transitional relationships between those states (e.g., successive locations along often-traveled trajectories). The hippocampal cognitive map is believed to be one such internal model, supporting a variety of behaviors, including associative learning, navigational planning, and inference. It remains unknown which facets of hippocampal coding are required for these different behaviors and how they support both associative and predictive memory functions.

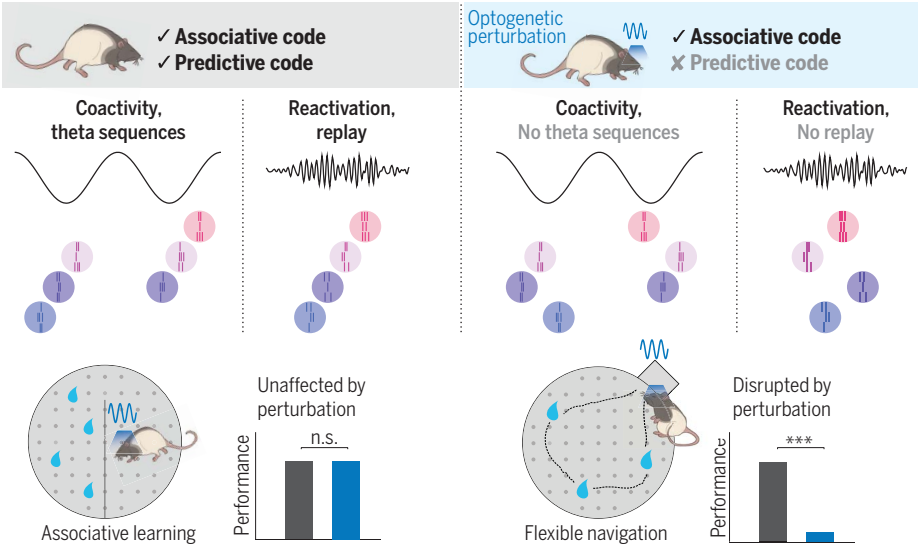
RATIONALE: We hypothesize that two modes of hippocampal activity support learning of world states and state transitions, respectively. On one hand, the synchronous coactivity of groups of hippocampal neurons—cell assemblies—may encode features of individual states, forming an associative code. On the other hand, the ordered activation of these cell assemblies into behaviorally relevant sequences may encode the relational structure between states, forming a predictive code. Previous research has not been able to dissociate these two dynamic codes or provide evidence of their specific functions. We leveraged an optogenetic approach to dissociate these two coding schemes, with the goal of disrupting the predictive code (hippocampal sequences) while preserving the associative code (rate coding and coactivity dynamics) in behaving rats. This dissociation allowed us to examine the different memory functions of these two codes.

RESULTS: We optogenetically perturbed the fine temporal coordination of hippocampal place cell firing as rats navigated specific spatial trajectories in a novel maze. This manipulation disrupted properties of the predictive code (such as temporally compressed place cell sequences and anticipatory place field shifts), but global network dynamics and single-cell spatial tuning and rate coding properties were preserved.

During sleep after the novel experience, we observed that task-related cell assemblies encoding discrete maze locations were reactivated in sharp wave–ripples (SWRs), unaffected by the manipulation. However, their sequential structure did not reproduce the order in which they were active in the task, resulting in impaired sequential replay for the perturbed trajectories. This result shows a dissociation between assembly reactivation and sequence replay, two phenomena previously assumed to reflect the same underlying process. The same manipulation did not disrupt replay of familiar trajectories, suggesting that the precise temporal coordination of place cell firing during learning mediates initial plasticity required for subsequent replay. Computational simulations suggest that distinct Hebbian plasticity mechanisms mediate assembly reactivation and sequence replay.

We tested the functional role of the predictive code by deploying our optogenetic manipulation in two different hippocampal-dependent memory tasks. Context-reward associative learning in a conditioned place preference task was unaffected and thus does not require a predictive map or memory replay. On the other hand, flexible memory–guided navigation in a foraging task was perturbed by the manipulation and thus depends on hippocampal predictive coding.

CONCLUSION: Our results provide a mechanistic and functional dissociation between coactivity and sequence codes in the hippocampus. Hippocampal cells with similar responses to behavioral variables fire together, forming functional assemblies during learning, which are reactivated in SWRs during subsequent sleep. These cell assemblies encode discrete states in the environment, an associative code that is sufficient for some types of episodic memories. As these cell assemblies are activated in a specific order during behavior, they form temporally compressed hippocampal sequences and promote Hebbian plasticity. This process enables the replay of behaviorally relevant sequences during SWRs. Hippocampal sequences thus encode transitional structures of world states, generating a predictive model on top of the associative code of individual assemblies. This new framework contributes to our understanding of how memory associations develop into predictive representations of the world and helps reconcile previously disparate views on hippocampal function. ■



Associative and predictive codes in the hippocampus. Our optogenetic manipulation perturbed hippocampal sequences without affecting cell coactivity, thus selectively disrupting the predictive code. After learning, cell assemblies were reactivated, but their order was not preserved, abolishing sequential replay (right). Perturbing the predictive code had no effect on associative learning (bottom left) but did disrupt the flexible learning of novel optimal trajectories on a maze (bottom right). [Rat illustrations: Yu Kang]

Department of Neurobiology and Behavior, Cornell University, Ithaca, NY, USA.
*Corresponding author. Email: afr77@cornell.edu
†These authors contributed equally to this work.
Cite this article as C. Liu *et al.*, *Science* **382**, eadi8237 (2023). DOI: 10.1126/science.adi8237

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.adi8237

RESEARCH ARTICLE

NEUROSCIENCE

Associative and predictive hippocampal codes support memory-guided behaviors

Can Liu†, Ralitsa Todorova†, Wenbo Tang, Azahara Oliva, Antonio Fernandez-Ruiz*

Episodic memory involves learning and recalling associations between items and their spatiotemporal context. Those memories can be further used to generate internal models of the world that enable predictions to be made. The mechanisms that support these associative and predictive aspects of memory are not yet understood. In this study, we used an optogenetic manipulation to perturb the sequential structure, but not global network dynamics, of place cells as rats traversed specific spatial trajectories. This perturbation abolished replay of those trajectories and the development of predictive representations, leading to impaired learning of new optimal trajectories during memory-guided navigation. However, place cell assembly reactivation and reward-context associative learning were unaffected. Our results show a mechanistic dissociation between two complementary hippocampal codes: an associative code (through coactivity) and a predictive code (through sequences).

Learning associations from experience and using this knowledge to make novel predictions are integral features of memory-guided behavior. Learning the structure of one's environment can be conceptualized as two separate processes: learning the states in the world and learning the likely transitions between those states. The hippocampus has been implicated in both processes because it has been shown to be fundamental for encoding spatial-temporal associations (*1*–*5*) and supporting flexible memory–guided behaviors, planning, and inference (*6*–*11*). Such involvement in diverse types of memory has led to different theories of hippocampal function focused either on its role in the formation and recall of episodic memories (*12*–*15*) or in the generation of predictive models to guide behavior (*10*, *16*, *17*), while some proposals highlighted the intrinsic relationship between these functions (*18*, *19*). However, it is unclear which facets of hippocampal coding are required for these different behaviors. One prominent candidate is the short-timescale (~10 to 30 ms) coincidental firing of hippocampal neurons with selective responses to external variables (e.g., place cells encoding the same location). This synchronous activation of "cell assemblies" can lead to the strengthening of their connections through Hebbian plasticity, forming a coactivity code (*20*–*22*). In addition, hippocampal cell assemblies organize into sequences. During navigation, temporally compressed neuronal sequences on the timescale of a single cycle of the theta oscillations (6 to 12 Hz; i.e., "theta sequences") sweep ahead of the animal's current position to encode future routes (*23*–*25*). During rest periods, the sequential activation of place cells "replays" past spatial trajectories (*26*, *27*). Coactivity and sequential hippocampal dynamics have been traditionally considered conjoined aspects of the same underlying process, supported by common circuit and synaptic mechanisms to the extent that they are regarded as nondissociable. An alternative possibility is that coactivity and sequential dynamics of hippocampal place cells represent distinct coding schemes (Fig. 1). First, assemblies of coactive cells would encode individual associations, or discrete states in the world. Second, assemblies can be concatenated into sequences reflecting either experienced or inferred behaviorally relevant state transitions, forming a predictive code. Although multiple studies have found deficits in memory-guided behavior by disrupting hippocampal activity (*28*–*32*), none have been able to dissociate these two dynamic modes and provide evidence of the specific role of hippocampal sequences in behavior. To overcome this limitation and test our hypothesis, we selectively disrupted hippocampal sequences in rats while preserving place cell expression and coactivity dynamics and investigated their specific contribution to predictive coding and flexible spatial learning.

Results

Disruption of hippocampal theta sequences impairs the formation of a predictive map

To dissociate assembly coactivation and the sequential organization of assembly activity, we sought to specifically disrupt place cell sequence dynamics while preserving their tuning and coactivity properties in the main hippocampal output region, CA1. Timing of CA1 place cell firing relies on inputs from the medial entorhinal cortex (MEC) (*33*–*35*). However, silencing or inactivating the MEC profoundly

Department of Neurobiology and Behavior, Cornell University, Ithaca, NY, USA.
*Corresponding author. Email: afr77@cornell.edu
†These authors contributed equally to this work.

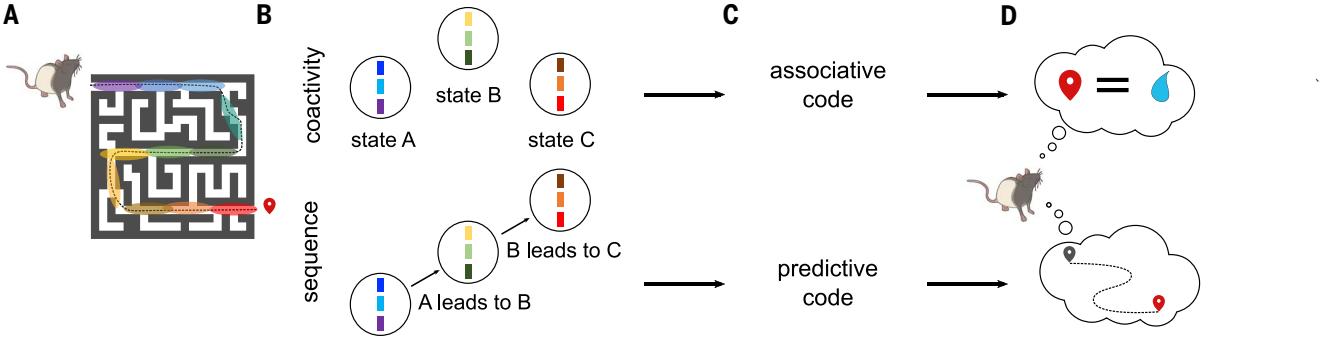


Fig. 1. Two hypothesized memory codes in the hippocampus. (**A**) As a rat learns to navigate a maze to find a reward, place cells activate along its spatial trajectory (color areas represent place fields). (**B**) During navigation, but also during offline periods, place cells with overlapping fields are active together, forming functional assemblies (top), and these assemblies form a sequence that reproduces an animal's learned trajectory (bottom). (**C**) We propose that the coactivity of cell assemblies represents an associative code to learn discrete states in the world, while their sequential activation forms a predictive code to learn behaviorally relevant state transitions. (**D**) An associative memory code may be sufficient for some types of memory, such as associating a place with a reward. Other types of memory-guided behavior, such as flexibly navigating a complex environment, would require a predictive model.

interferes with hippocampal physiology, affecting place cell expression, coactivity, and phase coding, and thus lacks the necessary specificity to test the role of sequences (*36*–*39*). To overcome these limitations, we expressed channelrhodopsin (ChR2) in rat MEC γ-aminobutyric acid–releasing (GABAergic) cells to optogenetically entrain them at an artificial gamma frequency (*40*, *41*). With this approach, we selectively perturbed the fine temporal structure of MEC outputs while simultaneously recording hippocampal population activity (Fig. 2A and fig. S1). In contrast to previous manipulations (*33*, *36*), our optogenetic perturbation spared theta power, frequency, and the theta phase modulation of CA1 unit firing (fig. S1). To evaluate the effect of our optogenetic perturbation on CA1 place cell properties, we delivered light while rats (n = 5) ran in one direction ("Stim ON") but not in the opposite direction ("Stim OFF") on a novel linear track. We aimed to disrupt specific sequences that encoded part of the experience and to leave other sequences intact as a control. The formation, stability, and firing properties of CA1 place cells at the single cell and population levels were preserved for both the Stim ON and Stim OFF directions (Fig. 2, B to F, and figs. S2 to S4). Place cells in the two directions had similar properties and formed a highly directional code such that Stim OFF and Stim ON trajectories could be decoded separately (fig. S3A), and on the population level, the composition, spatial properties, and stability of place cell assemblies were similar in the Stim OFF and Stim ON directions (figs. S4, E to H). Despite preserved place cell tuning curves, the progressive shift in their spike timing in relation to the phase of global theta rhythm ["phase precession" (*42*)] was impaired in the Stim ON direction (Fig. 2D and fig. S5). Moreover, the sequential firing of pairs of place cells at the theta timescale was disrupted in Stim ON (Fig. 2, G and H). In contrast, across theta cycles, cofiring of CA1 cells was preserved in both directions (Fig. 2, G and I).

In our framework, the formation of place fields (PFs) in a novel environment reflects learning the states in the world (e.g., locations), whereas the formation of a predictive map develops as an animal learns the likely transitions between those states (e.g., successive locations along often-traveled trajectories). To test whether the predictive code could be dissociated from the encoding of the individual states with our manipulation, we examined the evolution of place cells and hippocampal predictive coding features in the novel maze. We found that place cell properties in both Stim OFF and Stim ON directions showed experience-dependent development. Place field quality in both directions improved across laps, as measured by both their spatial information (Fig. 3B) and the decoding accuracy of animal position from place cell population activity (fig. S3), indicating that spatial maps were formed and developed with experience, unaffected by our optogenetic perturbation.

We then examined the development of a predictive code in two ways. We first analyzed theta sequences, because they are temporally compressed representations of spatial trajectories, reflecting short-term predictions of the immediate future (*43*–*46*). In contrast to the preserved spatial map formation, theta sequences were impaired in the Stim ON direction (Fig. 3A). Moreover, we observed a specific experience-dependent development of theta sequences in the Stim OFF direction, such that they became more prospective across laps (encoding longer distances ahead of the animal) (Fig. 3C), a process that could reflect the formation of increasingly predictive internal representations of the maze. This development was not observed in the Stim ON direction (Fig. 3C). Another property of the hippocampal predictive map is the progressive backward shift of place fields toward past states (or locations) with experience (*47*), so that the early firing reflects the expectation of upcoming states (*17*). We observed a progressive backward shift across laps of place fields only in the Stim OFF direction, as expected, not the Stim ON direction (Fig. 3, D and E). These changes in the predictive code were not due to differences in animal behavior between the two trajectories (fig. S3G).

Impaired replay with preserved reactivation of a novel experience

Memory formation in the hippocampus is typically measured as both the reactivation of short-timescale coactivity patterns ("assemblies") and the replay of neuronal sequences during sharp wave–ripples (SWRs) (*48*–*50*). We first examined sequential replay of a running experience on a novel linear track. During pauses between track running and during sleep after the task, we observed SWR-associated replay events reproducing place cell sequences of the Stim OFF direction (Fig. 4, A and B). However, replay sequences of the Stim ON trajectory, where hippocampal predictive coding was disrupted, were impaired (Fig. 4B). Compared with baseline sleep sessions before the task, the fraction of SWRs that expressed significant replay of the Stim OFF trajectory and their sequence scores were considerably increased, but no experience-dependent enhancement was observed for the Stim ON trajectory (Fig. 4C and fig. S6). This effect cannot be explained by differences in single neuronal activity during SWRs or decoding quality (fig. S7). Stim ON replay sequences were not observed in either the forward or the reverse direction (fig. S6, B and C), and this result was replicated with different replay methods and decoding parameters (figs. S6 and S7). We also tested whether this disruption of replay of specific trajectories was due to the impaired formation of a predictive map (i.e., encoding the relevant state transitions). We conducted an experiment in which rats (n = 3) first explored a linear track without manipulation to form a predictive map without perturbation and, in a subsequent session, experienced optogenetic perturbation in one direction as previously described. Theta sequences were still impaired in the Stim ON direction, but we observed intact replay for both Stim ON and Stim OFF trajectories (fig. S8), confirming that although predictive coding is key in the initial stages of encoding novel experiences, once the predictive map is established, it does not require further repetitions to generate replay of recent experience.

In the novel linear track experience, we next examined reactivation of task coactivity patterns by detecting synchronous cell assemblies that encoded discrete positions along both Stim OFF and Stim ON trajectories (fig. S9). The composition, spatial properties, and stability of assemblies were similar in the Stim OFF and Stim ON directions (fig. S4, E to H). Surprisingly, Stim OFF and Stim ON assemblies reactivated with similar strength during SWRs (Fig. 4, A and B), significantly increasing their activity compared with the preceding sleep (Fig. 4D and fig. S9, B to D). This result indicates an intact memory reactivation of locations along both trajectories, which was also confirmed using other canonical metrics of memory reactivation (Fig. 4E and fig. S9A). We wondered how these stark differences in reactivation and replay can coexist. A closer look at the temporal dynamics of cell assemblies revealed an important difference. Assemblies from the Stim OFF trajectory reactivated during SWRs in a temporal sequence that reproduced the order of their encoded spatial locations in the maze (Fig. 4 and fig. S9A). However, while individual assemblies for the Stim ON trajectory reactivated during SWRs with similar strength to those for the Stim OFF trajectory, their task sequential structure was not preserved in these reactivation events (Fig. 4B and fig. S9E), resulting in the impairment of sequence replay and the preservation of reactivation of the Stim ON trajectories. Indeed, although the strength of assembly reactivation and replay for the Stim OFF direction were correlated at the level of individual SWRs, this correlation was absent for the Stim ON direction (Fig. 4F), dissociating reactivation from replay in the absence of a predictive code.

Associative and predictive codes support different memory-guided behaviors

Next, we asked whether the associative (coactivity) and predictive (sequences) hippocampal codes have different functional roles. A predictive model of the environment would be required to learn tasks that necessitate flexible navigation. To test this prediction, we trained

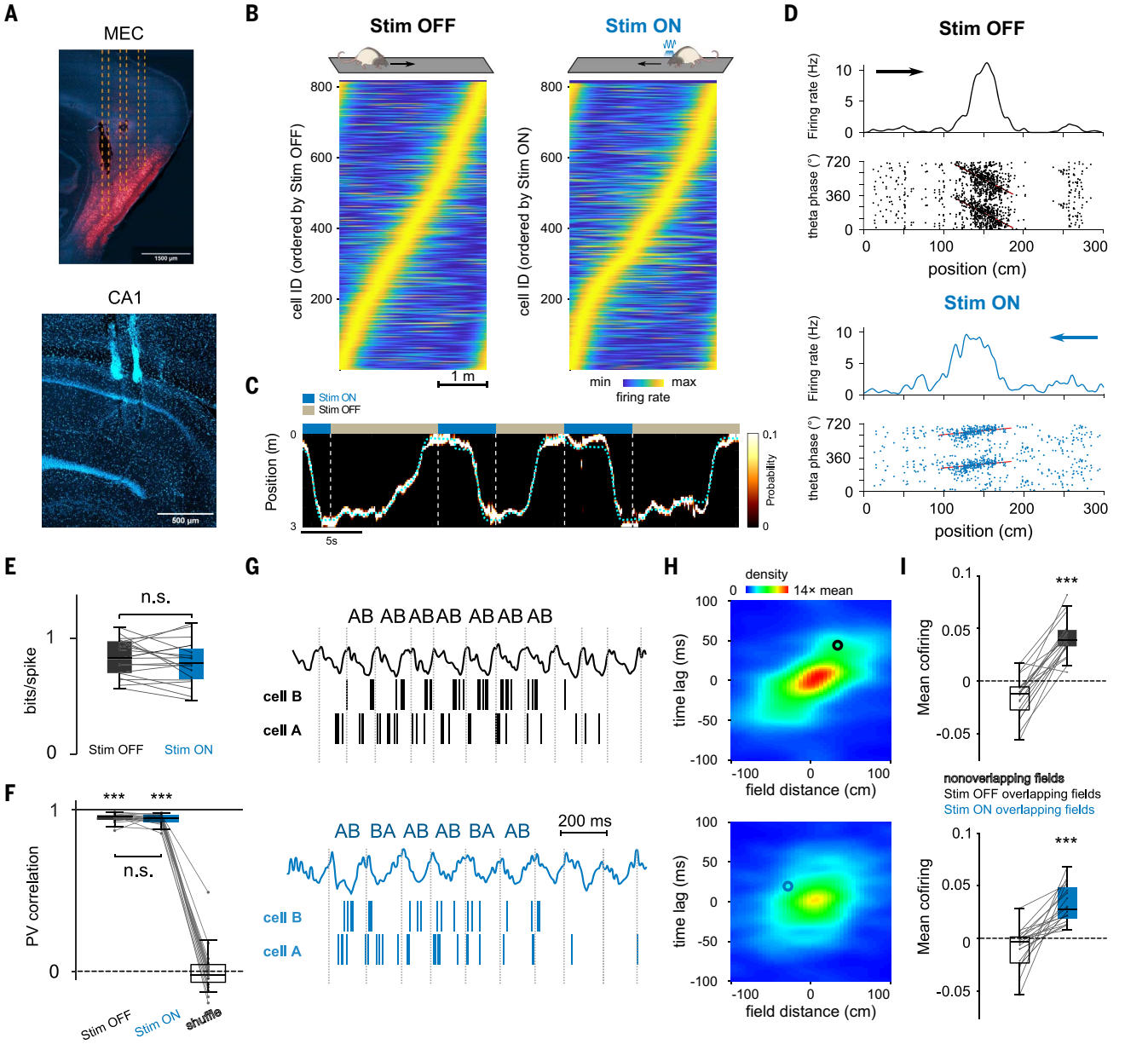


Fig. 2. Optogenetic perturbation of MEC gamma timing impairs temporal but not rate coding features of CA1 place cells. (**A**) Histological sections showing (top) expression of mDlx-ChR2-mCherry in MEC GABAergic cells (red) and optic fibers' tracks (dashed lines) and (bottom) probe locations in CA1. (**B**) Rate maps for CA1 pyramidal cells (n = 815) in the Stim OFF and Stim ON directions. (**C**) Example Bayesian decoding of CA1 population spike trains during running behavior. Dashed cyan line represents animal's actual position. The color map indicates probability of a decoded location. (**D**) Representative place cells for the Stim OFF (top) and Stim ON (bottom) directions. (Top) Firing rate as a function of spatial positions. (Bottom) Theta phase–position raster of place cell spikes. Red line represents the linear circular regression. (**E**) Spatial information was not significantly different between cells in the Stim OFF and Stim ON directions (P = 0.39; Wilcoxon signed-rank test). (**F**) Representation stability as measured by the population vector (PV) correlation between even and odd trials in both directions was higher than shuffled data (OFF/ON versus shuffle: $P = 6 \times 10^{-8}$/$P = 6 \times 10^{-8}$, Wilcoxon signed-rank test) and not different between the ON and the OFF directions (P = 0.35, Wilcoxon signed-rank test). (**G**) Examples of theta firing relationships for cell pairs with overlapping place fields in OFF (top) and ON (bottom) directions. The cells cofired in the same theta cycles in both cases, but the order of firing (AB) is only consistent in the OFF direction. (**H**) Average normalized density plot between place field distance and theta timescale firing lag for all overlapping place cell pairs in OFF (top) and ON (bottom) running directions. Theta compression was disrupted in Stim ON only ($P = 9 \times 10^{-5}$, Wilcoxon signed-rank test between theta compression slopes; see methods). Circles indicate the cell pairs shown in (G). (**I**) Cofiring of overlapping place cells in the same theta cycles was preserved in both directions (overlapping versus nonoverlapping place cells in OFF/ON: $P = 1 \times 10^{-4}$/$P = 9 \times 10^{-5}$, Wilcoxon signed-rank test; OFF versus ON: P = 0.067). ***$P < 0.001$.

rats (n = 5) to learn a novel spatial configuration of rewards in a familiar "cheeseboard" maze (*8*, *41*, *51*). Rats learned to find three hidden water rewards in different locations on the maze, with the reward locations changing every day (Fig. 5, A and B). Each trial started with the animal in the home box and ended when they retrieved the three rewards and came back to the box. The most efficient strategy for the animal in this task was thus to find the optimal route that connected all three reward locations by learning to sequentially

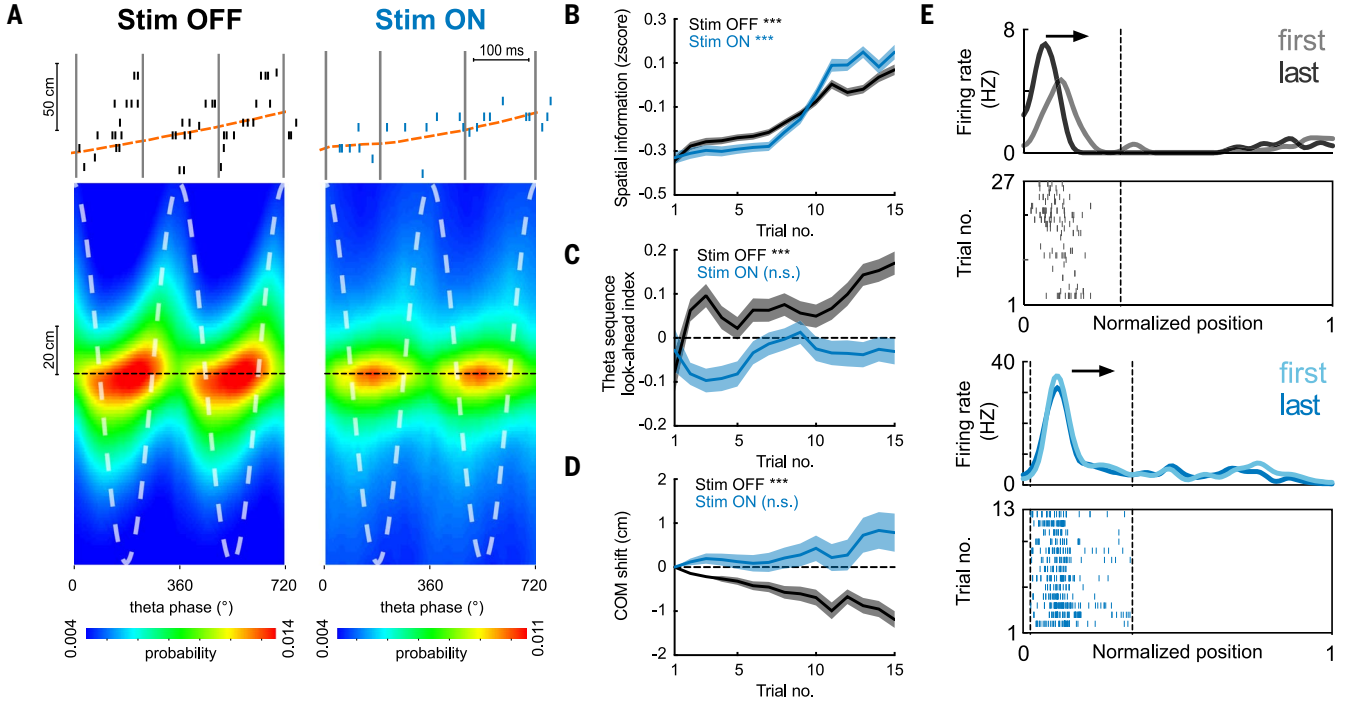


Fig. 3. Impaired development of place cell theta sequences and predictive properties. (**A**) (Top) Example theta sequences. (Bottom) Average (n = 20 sessions in five rats) decoded position estimated across different theta phases, relative to the actual animal position (dashed black line). Stim ON theta sequences were impaired (Stim OFF versus Stim ON quadrant scores: $P = 7 \times 10^{-141}$; weighted correlation: $P = 4 \times 10^{-22}$, Wilcoxon rank sum test, n = 17,992/13,265 theta cycles). (**B**) Increase in spatial information of place cells across laps in both Stim ON (blue) and Stim OFF (black) conditions (***$P = 1.8 \times 10^{-11}$/1.2×10^{-5} for Stim OFF/ON, Wilcoxon rank sum tests comparing the 1st versus the 15th lap). (**C**) Increase in theta sequence look-ahead index in the Stim OFF (***$P = 4.0 \times 10^{-6}$), but not the Stim ON, direction (P = 0.71, Wilcoxon signed-rank tests). (**D**) Backward shifting of PF center-of-mass (COM) overlaps on the Stim OFF trajectory (***$P = 1.3 \times 10^{-69}$) but not the Stim ON trajectory (P = 0.16, Wilcoxon signed-rank tests). COM shift relative to the first lap is shown (>0, forward shifting to future; < 0, backward shifting to past). (**E**) Example place cells showing backward shifting place fields on the Stim OFF trajectory (top) and stable fields on the Stim ON trajectory (bottom). Arrows depict the animal's running direction. Curves on top show smoothed firing rate for the first and last laps, and raster plots below show spikes across all laps. For all these analyses, n = 20 sessions from five rats were used.

navigate from one reward to the next. On half of the days, we performed the same optogenetic perturbation during learning trials (Stim ON sessions), whereas in the other half there was no stimulation (Stim OFF sessions). We observed an intact formation of place fields (fig. S10) and impaired theta sequences (Fig. 5D) in Stim ON sessions. Animal learning performance was impaired in Stim ON compared with Stim OFF sessions (Fig. 5C and fig. S10), supporting a role of the predictive code in the efficient learning of novel optimal trajectories in a familiar environment. We next asked whether replay would be preserved despite the disruption of theta sequences in this familiar environment. We first examined replay of the novel trajectory learned by the end of the current day's training session. In Stim OFF sessions, we found robust replay in both awake SWRs (during intertrial periods in the home box) and in the post-learning sleep session (Fig. 5F and fig. S11A). In contrast, replay of the novel route to the new goal locations was impaired in Stim ON sessions (Fig. 5F and fig. S11A), suggesting that when learning a new optimal route to goals, even in familiar environments, the transitional structure between locations of the route encoded by theta sequences is required for its subsequent replay. Replay of the familiar routes to previously learned locations was preserved in both Stim OFF and Stim ON sessions (fig. S11B). These effects were not due to differences in SWR properties, single-unit firing dynamics during SWRs, or decoding accuracy of animal position between Stim OFF and Stim ON sessions (figs. S11, D and E, and S10C). The two-dimensional (2D) place field properties of individual cells were also preserved, similar to those on the 1D linear track (fig. S10E).

In contrast to the impaired Stim ON replay, reactivation of cell assemblies was maintained in both Stim OFF and Stim ON sessions (Fig. 5G and fig. S11), replicating our results from the linear track. Finally, we tested whether impairment of hippocampal sequences with preserved place cell expression and coactivity affected subsequent memory recall. Memory recall measured 2 and 22 hours after learning was disrupted in Stim ON compared with Stim OFF sessions (Fig. 5H), and animals spent less time exploring around the reward locations after Stim ON than after Stim OFF sessions (fig. S10B). Furthermore, while performance on the 2- and 22-hour tests during Stim OFF sessions remained on par with the final performance reached during learning, in Stim ON sessions, performance was further deteriorated (fig. S10A).

Our hypothesis predicted that in contrast to "model-based" behaviors that require a predictive code, an associative memory code implemented by the coactivity of hippocampal cells would be sufficient to support other types of episodic memory, such as associating spatial context with rewards. We trained rats (n = 5) on a conditioned place preference (CPP) paradigm (Fig. 5I), which depends on intact hippocampal function. Rats were trained to associate one of the two halves of a cheeseboard maze with the presence of water rewards at random locations. After four pairing trials, we tested their preference for either of the two halves of the maze (without reward present). In half of the sessions, we delivered the optogenetic perturbation during all running periods during pairing trials. In line with our previous results, we detected similar assembly reactivation during sleep after both Stim OFF and Stim ON pairing

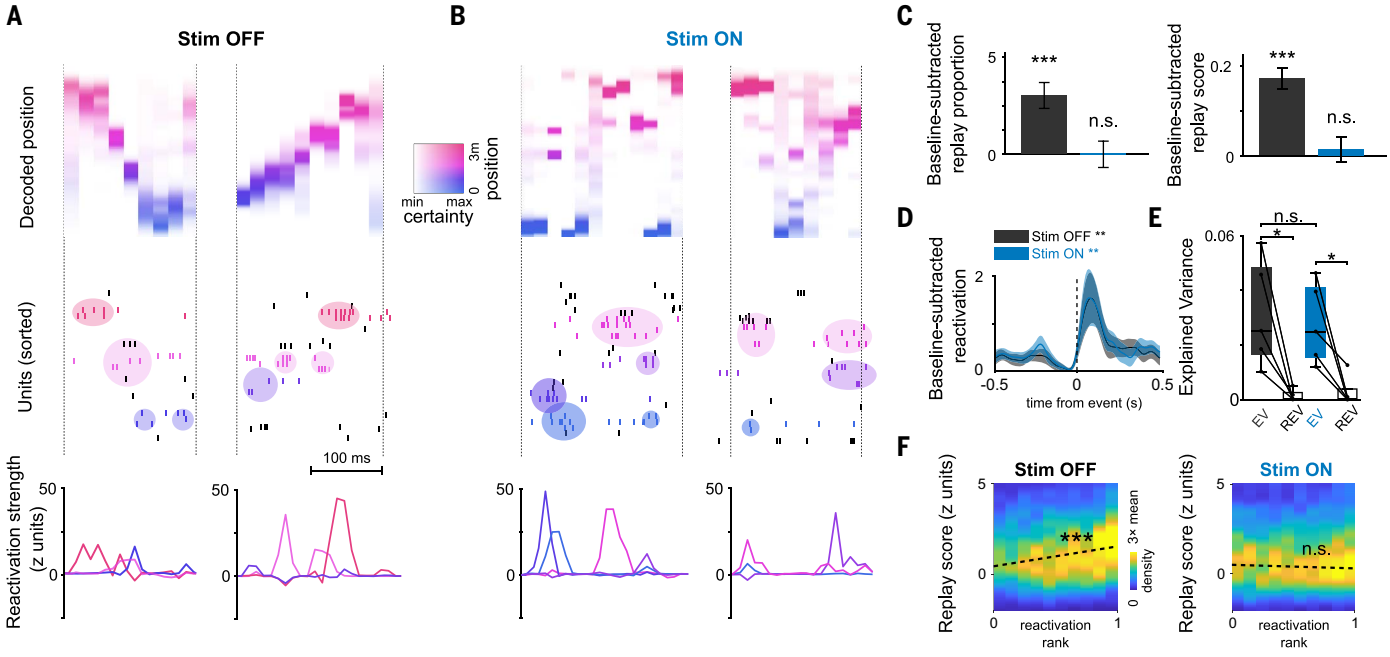


Fig. 4. Impaired replay with preserved reactivation of a novel experience. (**A** and **B**) Examples of replay and reactivation for (A) the Stim OFF and (B) the Stim ON experience, respectively. (Top) Decoded position during replay events (color coded by decoded position). (Middle) Raster plots of neuronal firing; colored circles indicate activations of assemblies composed of neurons with nearby place fields (colored ticks; colors of assemblies reflect their peak activation position). (Bottom) Assembly reactivation strength curves. (**C**) Increase in proportion significant replay (left) and sequence scores (right) of SWRs in post-task sleep as compared with pre-task sleep. Unlike the OFF direction, there was no significant replay in the ON direction (n = 4437 events; Stim OFF: student's t test, $^{***}P = 7 \times 10^{-11}$; Stim ON: P = 0.29 for proportions; and Wilcoxon signed-rank test: OFF $^{***}P = 3 \times 10^{-17}$ and ON P = 0.29 for scores). Only the first session in the maze was included (n = 5 sessions). (**D**) Reactivation strength of task-related assemblies centered on post-task sleep SWR, increased relative to baseline (Stim OFF post-task sleep versus baseline sleep during SWRs: Wilcoxon signed-rank test, n = 31 components, $^{**}P$ = 0.0096; Stim ON post-task sleep versus baseline sleep during SWRs: Wilcoxon signed-rank test, n = 27 components, $^{**}P$ = 0.0026). (**E**) Reactivation of pairwise neuronal correlation as measured by explained variance (EV) in post-task sleep was significant for both directions (Stim OFF: $^{*}P$ = 0.031; Stim ON: Wilcoxon signed-rank test, $^{*}P$ = 0.031; Stim OFF versus Stim ON, Wilcoxon signed-rank test, P = 0.81; n = 5 sessions). (**F**) Correlation between replay scores and assembly reactivation strength for post-task SWRs computed independently for Stim OFF and Stim ON replay events. Reactivation and replay were correlated in Stim OFF (Pearson's correlation coefficient r = 0.11, $^{***}P = 3 \times 10^{-12}$) but not in Stim ON (Pearson's r = −0.028, P = 0.06).

trials (fig. S12). Population activity during SWRs formed distinct memory representations for both contexts (rewarded and nonrewarded maze halves) in both Stim OFF and Stim ON sessions, consistent with an intact associative code (fig. S12). Because the CPP task involves boundaries, it gave us the opportunity to quantify a critical signature of hippocampal predictive coding, which is the elongation of place fields along obstacles in 2D mazes, reflecting the topological structure of the space (*17*, *52*). Similar to previous reports (*53*, *54*), fields near the boundary of our maze tended to elongate parallel to the wall in Stim OFF sessions (Fig. 5K and fig. S12). However, this effect was disrupted during Stim ON (Fig. 5K and fig. S12), providing further support that the predictive map was impaired. To assess replay in the absence of stereotypical trajectories, we examined the degree to which the sequential order of firing in post-task sleep SWRs resembled the sequential order in behavior. The sequential order was preserved in post-task sleep in Stim OFF sessions but not in Stim ON sessions (fig. S12, F and G). Despite these changes in the predictive code, on the test session, where no reward or stimulation was delivered, rats showed an enhanced preference for the previously rewarded context in both Stim OFF and Stim ON conditions (Fig. 5, J and L), confirming that an associative memory code was sufficient for the formation and retrieval of a context-reward association.

Circuit mechanisms for coactivity and sequence memory codes

Given this dissociation, we wondered which distinct circuit mechanisms support the reactivation and replay of memories. The main inputs that modulate the firing dynamics of CA1 pyramidal cells are excitatory projections from CA3 and the entorhinal cortex as well as perisomatic inhibition (Fig. 6A). The relative strength of each of these inputs during behavior can be estimated by measuring layer-specific gamma oscillations (*34*, *35*, *55*, *56*). MEC optogenetic perturbation specifically disrupted midfrequency gamma oscillations (gamma_M) in the distal CA1 dendrites, the domain innervated by MEC axons (Fig. 6B and fig. S13), while other CA1 gamma oscillations, elicited respectively by CA3 inputs and local inhibition (*34*, *35*, *55*–*57*), remained unaltered (Fig. 6B and fig. S13). The coupling of CA1 spikes to gamma_M, but not to the other two gamma oscillations, was also impaired (fig. S13), consistent with the notion that our manipulation disrupted CA1 pyramidal cell entrainment by MEC inputs but not by CA3 or local interneurons. The disruption of theta sequences during initial experience led to the impairment of replay but not the reactivation of assemblies, raising the question of which mechanisms link these processes. To answer this question, we turned to computational modeling.

Previous work has suggested that the highly recurrent CA3 hippocampal region with its auto-associative network structure is responsible for both the coactivation of CA1 cells forming functional assemblies (*22*, *58*) and the generation of sequential activity (*24*, *59*, *60*). We implemented a spiking CA3-CA1 network model capable of generating spontaneous replay. Our CA3 network consisted of recurrently connected leaky integrate-and-fire pyramidal

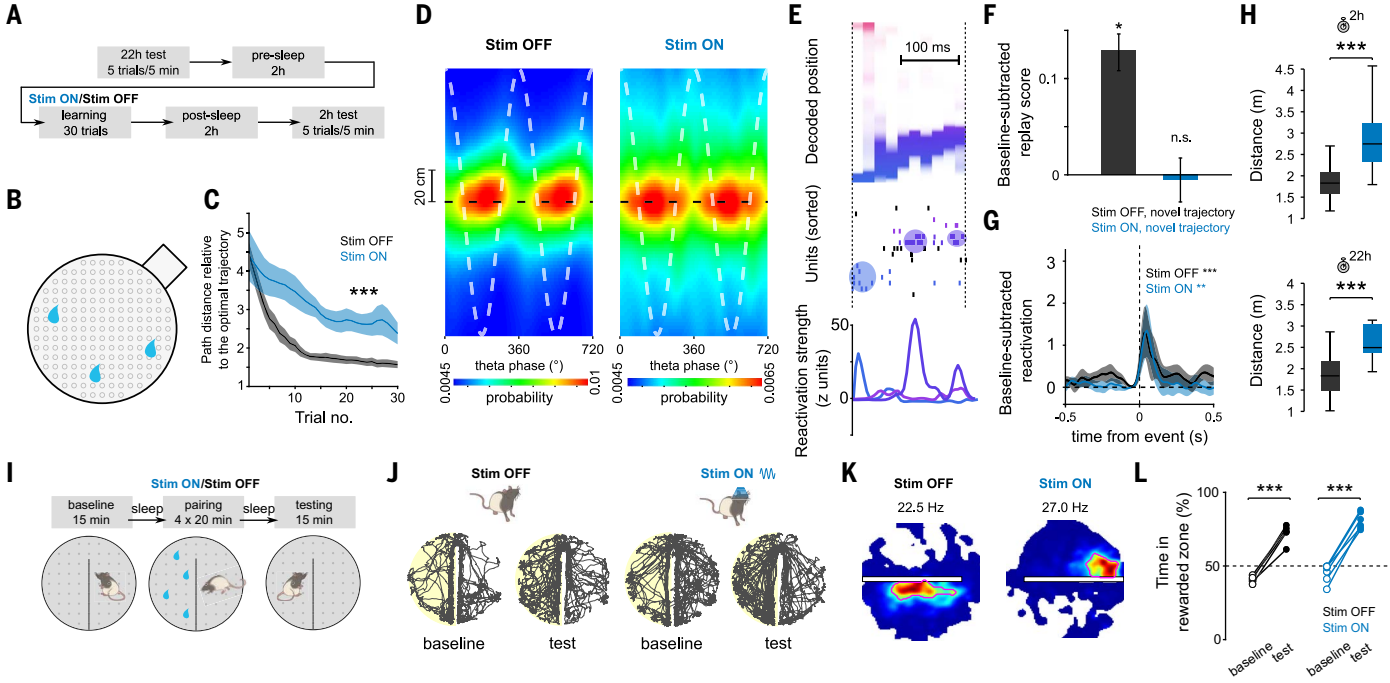


Fig. 5. Internally generated sequences are needed for memory-guided behavior. (**A**) Task structure of the cheeseboard task: Each day, animals learned a novel trajectory to get three hidden rewards in a circular arena. (**B**) Schematic of the cheeseboard setup. (**C**) Learning performance measured as distance traveled relative to optimal trajectory. ANOVA with repeated measures showed a significant main effect of group ($F_{1,43} = 133.4$, $P < 10^{-10}$; $n = 33$ and 11 sessions for Stim OFF and Stim ON, respectively). (**D**) Average decoded position versus theta phase, relative to actual animal position (dashed black line). Stim ON theta sequences were degraded (quadrant scores were lower in Stim ON than in Stim OFF: $n =$ 5958/2695 cycles for Stim OFF and Stim ON, respectively, $P = 2 \times 10^{-3}$; weighted correlations were lower in Stim ON than in Stim OFF: $P = 3 \times 10^{-4}$, rank sum test). Two cycles are shown for visibility. (**E**) Example of a decoded replay event. Decoded linearized position (top), spike raster (middle), and assembly reactivation strength (bottom) were shown for the same replay event. (**F**) Increase in replay score in post-task sleep as compared with pre-task sleep. Unlike the Stim OFF condition, there was no significant replay in the Stim ON condition (Stim OFF: $P = 3 \times 10^{-20}$; signed-rank test, $n = 11{,}460$ events; Stim ON: $P = 0.41$, $n = 8689$ events). (**G**) Reactivation strength of task-related assemblies centered on post-task sleep SWR (Stim OFF post-task sleep versus baseline sleep during SWRs: $n = 36$ components, $P = 4.71 \times 10^{-4}$; Stim ON post-task sleep versus baseline sleep during SWRs: $n = 22$ components, $P = 3.19 \times 10^{-3}$, signed-rank test). (**H**) Memory performance during 2 and 22 hour post-learning recall tests ($P = 6.8 \times 10^{-4}$/7.9×10^{-4} for 2 and 22 hour tests, rank sum test; $n = 25$/11 for Stim OFF/ON sessions from five rats). (**I**) Task structure of CPP task. (**J**) Example rat paths for Stim OFF (left) and Stim ON (right) baseline and testing sessions on CPP task. The side rewarded during pairing is highlighted in yellow. (**K**) Example place fields near the boundary for Stim OFF (left) and Stim ON (right) conditions, illustrating place field elongation along the boundary, which was disrupted in the Stim ON condition. Red circles emphasize the shape of the place field. (**L**) CPP memory performance: both Stim ON and Stim OFF training resulted in animals spending more active time in rewarded versus unrewarded side (Stim OFF: paired t test, $n = 5$ sessions, $P = 2.85 \times 10^{-4}$; Stim ON: paired t test, $n = 6$ sessions, $P = 2.10 \times 10^{-5}$).

cells and interneurons (*61*). This network was connected in a feedforward manner to a CA1 network, with a similar composition but lacking recurrent excitation, so every CA1 pyramidal cell integrated inputs from multiple CA3 pyramidal cells (Fig. 6C). We first simulated a learning phase, akin to a rat running on a novel linear track, in which place cells received random spatial inputs. In accordance with experimental observations, CA3-CA1 synaptic weights were updated according to an asymmetric (*62*, *63*) spike time–dependent synaptic plasticity (STDP) rule, whereas CA3-CA3 weights followed a symmetric (*64*) STDP rule (Fig. 6C). On the basis of our experimental findings, we simulated a Stim OFF condition, in which place cells displayed phase precession and robust theta sequences (Fig. 6D and fig. S14). In a Stim ON condition (mimicking the effects of our optogenetic MEC perturbation), place cells were phase-locked to theta oscillation but lacked phase precession and theta sequences (Fig. 6D and fig. S14), reproducing our experimental findings. We also simulated offline epochs [akin to non–rapid eye movement (non-REM) sleep] before and after spatial learning; in such periods, the network received low-level stochastic spiking inputs to drive activity. We examined CA1 spiking patterns after learning and quantified replay and reactivation as we did with our experimental data. In the Stim OFF condition, synapses between overlapping place cells were potentiated by STDP (fig. S14). Place cell sequences experienced in the learning phase were spontaneously replayed in the following offline epoch (Fig. 6, E and F). In the Stim ON condition, disorganized timing within theta cycles disrupted CA3-CA1 STDP, which relies on a consistent timing between pre- and postsynaptic spike pairs (fig. S14). CA3-CA3 STDP was preserved due to place cell coactivity within theta cycles, regardless of their precise temporal ordering (fig. S14). These effects resulted in an impairment of replay in the Stim ON condition (Fig. 6, E and F) but robust reactivation of place cell assemblies in both conditions (Fig. 6, E and G), mirroring our experimental results. The dissociation between reactivation and replay in our model suggests that reactivation and replay are differently mediated by STDP at CA3 and CA1 synapses, providing a plausible circuit-level mechanism for our findings.

Discussion

In this study, we described two complementary hippocampal circuit mechanisms that support the formation of memory associations and the generation of predictive representations of the environment. First, the coactivity of place cell assemblies encoded discrete states in the environment and, with their subsequent reactivation,

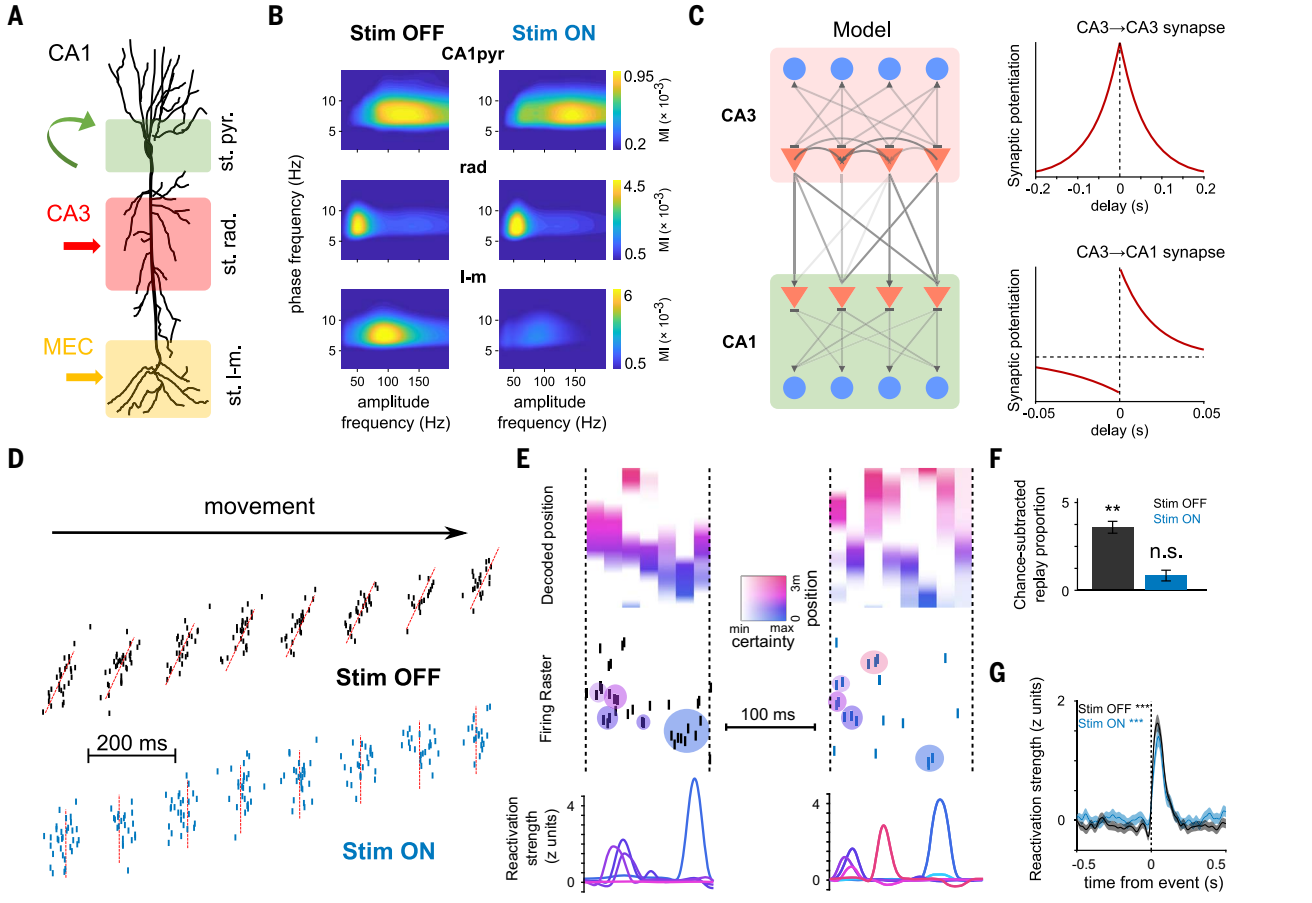


Fig. 6. Circuit mechanisms for coactivity and sequence hippocampal dynamics. (**A**) Schema depicting the inputs to CA1 pyramidal neurons stratified along their somatodendritic axis. Local inhibitory inputs are dominant in the pyramidal layer (st. pyr.), CA3 inputs target proximal apical dendrites in the stratum radiatum (st. rad.), and entorhinal inputs target distal dendrites in the stratum lacunosum-moleculare (st. l-m.). (**B**) Gamma amplitude-theta phase comodulograms for each layer-specific LFP component (see methods) displayed modulation in a specific gamma sub-band (averaged data from n = 10 sessions from four rats): CA1pyr in gamma$_F$ (100 to 160 Hz), rad in gamma$_S$ (30 to 60 Hz), and LM in gamma$_M$ (60 to 110 Hz). (Right) MEC perturbation selectively impaired LM gamma$_M$ but not CA1pyr or rad oscillations (**P = 0.002, Wilcoxon signed-rank test; n = 13 sessions from five rats). (**C**) (Left) Model schematic depicting a subnetwork of CA3 cells projecting to a subnetwork of CA1 cells. Triangles represent pyramidal neurons, and circles represent inhibitory interneurons. (Right) STDP rules used to train different synapses within the network during learning trials. (**D**) In Stim OFF simulations (top), place cells displayed phase precession and prominent theta sequences. In Stim ON simulations (bottom), phase precession was disrupted, and theta sequences were abolished. (**E**) Example replay events simulated by the model after Stim OFF (left) and Stim ON (right) learning. Decoded position, spike raster, and assembly reactivation strength as in Fig. 4. (**F**) Increase in proportion of "SWR" events with significant replay in post-task "sleep" epochs in Stim OFF and Stim ON protocols. Replay improvement was above chance levels in Stim OFF but not Stim ON simulations (Stim OFF: P = 0.0039, Wilcoxon signed-rank test, n = 6585 events; Stim ON: P = 0.19, n = 6031 events). (**G**) Reactivation strength of task-related assemblies centered on post-task sleep SWR-like events increase relative to baseline (Stim OFF post-task sleep versus baseline sleep during SWR-like events: Wilcoxon signed-rank test, n = 362 components, ***$P = 5 \times 10^{-44}$; Stim ON post-task sleep versus baseline sleep during SWR-like events: Wilcoxon signed-rank test, n = 312 components, ***$P = 1 \times 10^{-27}$).

supported associative memory. Second, temporally compressed hippocampal sequences encoded transitional structures of states in the environment, supporting the formation of a predictive map. Replay of these sequences provided a mechanism to update and exploit this predictive map, but only after intact encoding of transitional structures of task states during learning. Disruption of trajectory-specific theta sequences impaired their subsequent replay but did not affect the reactivation of neuronal assemblies representing those same locations. Reactivation and sequence replay are therefore dissociable neural processes. The dependence of replay on learning environment transitional structures through theta sequences was highly specific for both space and time: Neither disruption of theta sequences for the same trajectory traveled in the opposite direction nor disruption of previously experienced trajectories affected replay. Furthermore, these results indicate that encoding the specific transitions between place cells along behaviorally relevant trajectories, rather than place field formation per se or global theta-timescale dynamics alone, is an instrumental mechanism for replay, explaining and extending previous observations on the relationship between theta population dynamics and replay (*65*–*67*). The disruption of temporally compressed hippocampal sequences impaired flexible memory–guided navigation but did not affect the formation and recall of contextual associations.

Our results advance a framework to unify previously disparate views of hippocampal function, including encoding cognitive maps for spatial and episodic memory (*68*–*70*), predictive maps for flexible navigation and planning (*10*, *16*, *17*), and multimodal memory representations for relational processing (*18*, *71*). We argue that the hippocampal role in encoding episodic memories relies on two complementary

mechanisms. First, the synchronous coactivity of cell assemblies binding different features of experience into cohesive representations enables the fast encoding of discrete states in the world and supports associative memory. Second, the formation of a predictive map is supported by internally generated hippocampal sequences, binding discrete successor states into a relational structure. Moreover, our results can account for two independent lines of evidence supporting previous theories of hippocampal predictive coding: predictive sequences and "successor representation" place fields. Our manipulation impaired the development of theta and replay sequences. Previous studies have shown that both theta and replay sequences not only reproduce animal recent experiences but can also construct novel predictive representations. Theta sequences encode trajectories ahead of the animal's current location and can represent alternative paths at decision points (*44*–*46*). Replay sequences can reproduce never-experienced paths (*72*, *73*) and dynamically change to reflect learned task contingencies (*11*, *74*–*76*) or represent available paths not yet taken (*77*, *78*). In both cases, internally generated hippocampal sequences could be conceptualized as a sampling process between a series of possible future states from a probabilistic generative model, and they were both impaired by our manipulation. In addition, experience-dependent backward expansion and deformation around obstacles of place fields are believed to reflect the encoding of a predictive map over the locations the animal expects to occupy in the near future (*17*, *52*). These features were disrupted by our manipulation, in contrast to place field formation and stability, which may be supported by different plasticity mechanisms (*79*). Together, these effects of our manipulation suggest that all these "generative" features of hippocampal representations are linked by common circuit mechanisms.

Although numerous processes could contribute to the generation of replay sequences (*80*–*83*), our results suggest that initial theta sequence–mediated STDP is necessary for the replay of newly learned spatial trajectories but not for memory reactivation. These results suggest a hierarchical organization of hippocampal assembly and sequence dynamics. Previous work investigating the postnatal development of hippocampal dynamics supports this dissociation. Synchronous cell assemblies are already present around postnatal day 17 in rats, with their reactivation in SWRs reflecting encoding of individual locations on a maze (*84*). However, theta sequences and replay encoding animal spatial trajectories appear only after postnatal day 21 (*66*, *84*).

The present findings agree with previous studies that investigated the contribution of the temporal coordinatization of hippocampal spike timing to memory. Pharmacological disruption of place cell theta rhythmicity and coordination, while preserving their rate coding properties, impaired spatial memory (*85*, *86*). In addition, disruption of the behavioral timescale (on the order of seconds) sequential structure of CA1 firing responses during temporal delays also impaired memory (*31*, *87*). These studies highlight the importance of the precise temporal coordination of hippocampal spike timing for memory.

Additional research is still needed to elucidate the precise contribution of different input pathways and cell types in the entorhinal-hippocampal network to temporal coding and sequence generation. MEC inputs reach CA1 via direct projections from layer III and also indirectly via layer II inputs to the dentate gyrus, CA3, and CA2 (*88*). Our manipulation was not restricted to either of these input pathways and therefore cannot precisely determine their contributions to the generation of CA1 sequences. Furthermore, a small subset of MEC GABAergic cells send direct long-range projections to CA1 (*89*). It is possible that our manipulation also affected these projections; however, their sparsity and the fact that the stronger GABAergic projections to CA1 arise from the lateral rather than the medial entorhinal cortex (*90*) limit their potential contribution to the results described here.

Overall, this study suggests the coexistence of complementary associative and predictive codes in the hippocampus. This dual code could account for the wide range of behavioral functions attributed to this brain structure in learning, memory, navigation, and planning.

Materials and methods

Surgical procedures

Rats (adult male Long-Evans, 300 to 500 g, 3 to 6 months old) were kept in the vivarium on a 12 hour light/dark cycle and were housed two per cage before surgery and individually after it. All experiments conformed to guidelines established by the National Institutes of Health and have been approved by the Cornell University Institutional Animal Care and Use Committee.

Silicon probe implantation was performed as described previously (*41*, *51*, *91*). Animals were anesthetized with isoflurane anesthesia and craniotomies were carried out under stereotaxic guidance. Silicon probes (NeuroNexus, Cambridge Neurotech, or Diagnostic Biosignals) were mounted on custom-made 3D-printed microdrives to allow precise adjustment of the vertical position of sites after implantation. The probes were inserted above the target region. Craniotomies were sealed with sterile wax. Two stainless steel screws were placed bilaterally over the cerebellum to serve as ground and reference electrodes. Several additional screws were driven into the skull and covered with dental cement to strengthen the implant. Finally, a copper mesh mounted on a 3D-printed resin base was attached to the skull with dental cement and connected to the ground screws to act as a Faraday cage, attenuating the contamination of the recordings by environmental electric noise and protecting the headgear. Three doses of analgesic were administered, with the first dose administered prior to surgery in order to cover 72 hours total. After post-surgery recovery, probes were moved gradually in 50- to 150-μm steps per day until the desired position was reached. Hippocampal layers were identified physiologically by unit activity and characteristic local field potential (LFP) patterns (*34*, *92*).

A variety of different silicon probes were implanted in the dorsal hippocampus [−4.0 to 4.5 mm anteroposterior (AP) from Bregma and 2.6 mm from midline (ML)]. Data from some of the animals included in this study have also been included in a previous study, and surgical procedures have been described in more detail there (*41*).

For optogenetic experiments, rats were also implanted with custom-made optic fiber arrays (three 200-μm core multimode fibers each, ~500 μm apart, connected to a single 2.5-mm steel ferrule; Doric Lenses) in both hemispheres over the MEC (−7.7/−8.4/−9.1 AP; ± 4.6 ML and 4.7/ 4.3/3.2 mm from the surface of the brain, for each fiber respectively).

Optogenetic experiments

For optogenetic experiments, rats were injected with custom-prepared AAV5-mDlx-hChR2(H134R)-mCherry from AddGene [plasmids were a gift from Dr. Gord Fishell (*40*)]. Three injections per hemisphere were performed along the dorsoventral MEC as follows: (i) −7.7 AP, ±4.6 ML, 4.7 mm depth, 200 nl; (ii) −8.4 AP, ±4.6 ML, 4.3 mm depth, 400 nl; (iii) −9.1 AP, ±4.6 ML, 3.2 depth, 700 nl. After injection, craniotomies were sealed, and animals recovered in the vivarium for 3 weeks. After this period, a second surgical procedure for implanting optic fibers and electrodes was performed, as described above. Optic fiber arrays were implanted in the same craniotomies performed previously for virus injection. For optogenetic stimulation, fiber array ferrules were connected with mating sleeves to 450-nm blue light-emitting laser diodes coupled to 2.5mm steel ferrules (PL-450, Osram).

Optogenetic perturbations were performed by delivering blue light, modulated with a positive 53 Hz current sinusoid using an isolated current driver (Thorlabs). Light intensity was calibrated for each animal during home cage recordings by analyzing the suppression of LFP gamma power in the stratum lacunosum-moleculare during stimulation. A minimum and maximum power of 3 and 6 mW, respectively, was used. In the linear maze task, light

stimulation was triggered when the animal crossed an infrared sensor near the end of the track (but before the reward port located at the very end of it) and stopped when it crossed a similar sensor at the other end, resulting in stimulation only during running periods. In the cheeseboard maze and CPP tasks, the same optogenetic stimulation was applied during all running periods in the maze during learning trials.

Behavioral and electrophysiological recordings

After surgery, animals were handled daily and accommodated to the experimenter, recording room, and cables for 1 week before the start of the experiments. Before the start of the behavioral experiment, the animals were water-restricted. Electrophysiological recordings were conducted using Intan RHD2000 interface board or Recording Controller (IntanTech) and 64-channel digital headstages (IntanTech or Diagnostic Biochips), sampled at 20 kHz. For all behavior tasks, animal position was recorded with an overhead camera (Basler) and tracked with DeepLabCut (*93*) and custom codes.

Linear track task

In the linear track, rats were trained to run back and forth to collect small sugar water rewards at both ends. Animals typically performed between 30 and 80 trials per day. The session was terminated when the animal was satiated, typically after 40 min. The linear track was placed 1 meter above the floor and was 300 cm long and 7 cm wide with 5- to 10-cm-high walls. Water rewards were automatically delivered in reward wells at both ends of the track. Baseline and post-task sleep were recorded in the home cage before and after the task, respectively.

Cheeseboard maze task

The cheeseboard maze was a circular platform (120 cm diameter), where the animals learned to find three goal wells that contained water rewards. A trial was completed once the animal had retrieved all rewards and returned to the start box to collect an additional food pellet reward. The locations of the goal wells changed daily but were fixed during a given session. This strategy required the animals to update their memory for the new goal locations in an otherwise familiar environment during each session. Note that between trials, there was always a delay of ~30 s. If the rat could not find all three rewards within 2 min, the trial was terminated, and the animal was returned to the start box. A pre-task test session of five trials (with the same configuration as the previous day) was run every day to assess whether the animal remembered the previous day's positions, followed by a 2-hour sleep session. Next, a 30-trial learning session was conducted (with or without light stimulation). After a 2-hour sleep session after the learning trials, a five-trial post-task test session (with the newly learned reward configuration) was also conducted to examine whether the animal remembered the newly learned locations. Learning performance was evaluated as the distance traveled from the start box to collect the three rewards and divided by the optimal trajectory (shortest possible path). A value of 1 then indicates an optimal path taken by the animal. Performance in the post-task test sessions was computed in the same way (Fig. 5H), and additionally it was compared directly with the learning performance on the final learning trial (fig. S10A). An additional measure of performance was computed as the proportion of exploration time spent within 5 cm the correct wells (fig. S10B). Stim OFF and Stim ON sessions were alternated in a counterbalanced manner across animals. The inclusion of data for the analysis required that the animal be pretrained for a week in this task.

Conditioned place preference task

The CPP task was performed in the same maze as in the cheeseboard maze task but separated into two compartments by a 30-cm-high wall in the midline. Animals had free access to both compartments connected by a common corridor at one end of the maze (Fig. 5I). The task included three stages spanning 3 days: baseline, pairing, and testing. For the baseline stage on day 1, animals were allowed to freely explore the two compartments for 15 min with no reward. A baseline place preference was then measured as the difference in active exploratory time spent in the two compartments. The compartment that animals explored less was selected as the rewarded side for the subsequent pairing sessions. Pairing sessions included one session after the baseline session on day 1, two sessions on day 2, and one session before the testing session (see below) on day 3. During the pairing sessions, animals explored the maze for 20 min, with hidden water reward delivered at random locations on the rewarded side. On day 3, after the final pairing session, a testing session was performed, during which animals again explored the maze without reward for 15 min and the place preference was measured as was done during the baseline session. Optogenetic stimulation was delivered when animals actively explored the maze during pairing sessions. Each CPP session was flanked by sleep recordings (around 2 hours) in the home cage. Learning performance of the CPP task was measured by comparing the place preference during the baseline versus the testing session (Fig. 5L).

Tissue processing and immunohistochemistry

After the termination of the experiments, animals were deeply anesthetized and perfused transcardially, first with 0.9% saline solution followed by 4% paraformaldehyde solution. The brains were sectioned into 70-μm-thick slices (Leica Vibratome). The sections were washed and mounted on glass slides with a fluorescence medium [Fluoroshield with DAPI (4′,6-diamidino-2-phenylindole), product no. F6057, Sigma, USA]. A confocal microscope (Zeiss LSM 800) was used to obtain high-quality photos.

Computational model

Architecture

We built a simplified network model consisting of 1250 pyramidal cells (PYR) and 100 inhibitory interneurons (INT) in area CA3 and 1250 pyramidal cells and 100 inhibitory interneurons in area CA1 using the Brian2 simulation environment (*94*). Each neuron was modeled as an adaptive exponential leaky integrate-and-fire unit with cellular adaptation (*61*). Briefly,

$$C_m \frac{dV(t)}{dt} = -\Big(g_L(V(t) - V_{rest}) - g_L \Delta T \exp\left(\frac{V(t) - \theta}{\Delta T}\right) + I_{syn}(t) + w(t)\Big)$$

where $V(t)$ is the membrane potential, C_m is the membrane capacitance, g_L is the leak conductance, V_{rest} is the reversal potential of the linear leak current, θ is the spike threshold, ΔT is the threshold sharpness, I_{syn} is the synaptic current, and $w(t)$ is the adaptation current, described by

$$\tau_w \frac{dw(t)}{dt} = a(V(t) - V_{rest}) - w(t)$$

where the parameter a describes the strength of the subthreshold adaptation. The synaptic current I_{syn} was computed as

$$I_{syn}(t) = g_{AMPA}(t)(V(t) - E_{exc}) + g_{GABA}(t)(V(t) - E_{inh})$$

where E_{exc} = 0 mV and E_{inh} = −70 mV are the reversal potentials of excitatory and inhibitory currents, respectively. Synapses were modeled as conductances with biexponential kinetics (*61*).

Synapses

Within each area, PYR→INT connectivity occurred at 10% probability, and INT→INT and INT→ PYR connectivity occurred at 25% probability. Whereas PYR→PYR connectivity was absent from CA1, CA3 PYR cells projected at 10% probability to PYR from both CA3 and CA1, modeling recurrent CA3 connectivity and CA3 inputs into CA1, respectively. Synaptic parameters were used as previously reported (*61*), with the modification that synaptic weights were initialized to follow a lognormal distribution $\exp(N(0,1)) \times w_{init}$, where $w_{init}^{CA3\to CA3} = 0.3$ and where $w_{init}^{CA3\to CA1} = 0.7$. To simulate learning as a result of the activity during exploration, STDP

was modeled for PYR→PYR synapses alone, where synaptic weights are updated according to an additive pair-based learning rule as follows

$$\Delta w_+ = A_+ \exp\left(-\frac{\Delta t}{\tau_+}\right) \text{ at } t_{post} \text{ if } t_{pre} < t_{post}$$

$$\Delta w_- = A_- \exp\left(-\frac{\Delta t}{\tau_-}\right) \text{ at } t_{pre} \text{ if } t_{pre} > t_{post}$$

CA3→CA3 STDP followed a symmetric rule, where $A_+ = A_- = 80$ pA, and $\tau_+ = \tau_- = 62.5$ ms, while CA3→CA1 STDP followed an asymmetric rule, where $A_+ = 800$ pA, $A_- = -A_+ \times 0.4$, $\tau_+ = 20$ ms, and $\tau_- = 40$ ms. All weights were cropped at $w_{max} = 40$ nS.

Exploration phase

Spike trains of PYR cells during exploration were generated as previously described (*61*). Briefly, 10% of all pyramidal cells in each region were designated as place cells, and place fields were randomly distributed along a 3-m-long simulated linear track. Exploration was simulated as 10 min of exploration time while the rat ran along the linear track at 50 cm/s. Silent cells (nonplace cells) fired at 0.01 Hz, and place cell spikes were sampled from a Poisson process at 20 Hz, with the sampling procedure ensuring an inhomogeneous Poisson process with a time-dependent rate $\lambda(t)$, which was the product of a Gaussian tuning curve (representing the neuron's place field) with width $\sigma = 7$ cm (yielding $l = 30$ cm place fields) and a theta component. The theta component was either a phase precession component to simulate the sequences of the Stim OFF condition or a phase-locking component to simulate the lack of sequences but preserved theta-cycle coactivity in the Stim ON condition. The Stim OFF theta component was

$$\theta^{OFF}(t) = \cos\left(2\pi f_\theta t + \frac{\pi}{2\pi \cdot Q}\left(x(t) - x^i\right)\right)$$

where Q represents the proportion of the linear track that the place field spans and takes numeric value of 30-cm field over 300-cm linear track or 0.1, and x^i is the position of the place field of cell i.

The Stim ON theta component was a phase-locking component designed to yield the same phase-locking value as the other condition's theta precessing component but abolishing sequential order of spikes

$$\theta^{ON}(t) = \cos\left(2\pi f_\theta t + \frac{\pi}{2\pi \cdot Q} N\left(\frac{\pi}{2}, \frac{\sigma}{l}\right)\right)$$

where l is the length of the track (3 m), and $N\left(\frac{\pi}{2}, \sigma\right)$ is a normal distribution with mean $\frac{\pi}{2}$ and standard deviation $\frac{\sigma}{l}$.

Offline phase

Before and after exploration-triggered STDP, the network was simulated using Brian2 to respectively model pre (baseline) and post "sleep" sessions. To drive the network, each CA3 pyramidal cell received random inputs in the form of uncorrelated Poisson spike trains with a pooled mean rate of 15 Hz. Under these conditions, the network spontaneously generated population burst events (akin to SWRs).

Analysis

All the modeling steps described above were performed n =10 times with different randomly generated initial conditions determining place field locations, initial weights, etc. This resulted in $n = 10$ simulated sessions, in which theta sequences, replay, and reactivation were analyzed applying the same methodology and parameters as we did to analyze experimental data described above.

Spike sorting and unit classification

Spike sorting was performed semiautomatically with KiloSort (*95*) (https://github.com/cortex-lab/; KiloSort), followed by manual curation using the software Phy (https://github.com/cortex-lab/phy) and custom designed plug-ins (https://github.com/petersenpeter/phy-plugins) to obtain well-isolated single units. Cluster quality was assessed by manual inspection of waveforms and autocorrelograms, and by the isolation distance metric. Multiunit, noise clusters, or poorly isolated units were discarded for analysis. Well-isolated units were classified into putative cell types using the Matlab package CellExplorer (*96*) (https://github.com/petersenpeter/CellExplorer). Spiking characteristics, including the autocorrelograms, spike waveforms, and putative monosynaptic connections derived from short-term cross-correlograms, were used to select and characterize well-isolated units. Three cell types were assigned: putative pyramidal cells, narrow waveform interneurons, and wide waveform interneurons. Two key metrics used for this separation were burst index and trough-to-peak latency. Burst index was determined by calculating the average number of spikes in the 3- to 5-ms bins of the spike autocorrelogram divided by the average number of spikes in the 200- to 300-ms bins. To calculate the trough-to-peak latency, the average waveforms were taken from the recording site with the maximum amplitude for the averaged waveforms of a given unit.

Detection of brain states

State scoring was performed as previously described (*41*, *51*). Briefly, the LFP was extracted from wide-band data by lowpass filtering (sync filter with a 450 Hz cutoff band) and downsampling to 1250 Hz. Broadband LFP, narrow-band theta frequency LFP, and estimated electromyogram (EMG) were used for state scoring. Spectrograms were computed from broadband LFP with a fast Fourier transform in 10-s sliding windows (at 1 s), and a principal components analysis was computed after a Z-transform. The first principal component reflected power in the low (<20 Hz) frequency range, with oppositely weighted power at higher (>32 Hz) frequencies. Theta dominance was quantified as the ratio of powers in the 5 to 10 Hz and 2 to 16 Hz frequency bands. EMG was estimated as the zero-lag correlation between filtered (300 to 600 Hz) signals across recording sites (*55*). Soft sticky thresholds on these metrics were used to identify states. High LFP principal component 1 and the low EMG were considered non-REM, the high theta and low EMG were considered REM, and the remaining data were taken to reflect the waking state.

For analysis of neural activity during active behavior (e.g., place cell and theta sequence analysis), only periods in which the animals run faster than 5 cm/s were included.

SWR detection

To detect SWRs, one channel around the CA1 pyramidal layer was chosen for the ripple detection, and one channel from CA1 stratum radiatum was chosen for sharp wave detection. The difference between the two channels was used as the basis for SWR detection. This difference signal was filtered with a low-pass filter at 55 Hz, and then local minima were detected as candidate events. The corresponding ripple power (amplitude of the difference signal in the bandpass 80 to 250 Hz) for each candidate event was recorded. A true event was considered when both a sharp wave and ripple oscillation were detected in the same window. K-means clustering was used to define clustering of SWR from non-SWR events, and manual curation was used to better define the boundary and remove outliers. The events were then expanded until the (nonclipped) ripple power fell below 1 SD; short events (<15 ms) and longer events (>400 ms) were discarded.

To refine the detection of SWR start and end points for reactivation and replay analyses (in order to avoid including empty bins), we detected population burst events as periods when the instantaneous population firing rate (binned in 1-ms bins and smoothed with a Gaussian kernel with 10-ms width) reached a peak of >2 SD and remained greater than the mean. The SWR start and end time points were set as the start and end time points of the population burst within the respective SWR, and SWRs without population bursts were discarded.

Place cell analysis

Spiking data and the tracked animal's position were binned into 3-cm-wide segments of

the camera field projected onto the maze floor, generating raw maps of spike counts and occupancy. A Gaussian kernel (SD = 3 cm) was applied to both raw maps of spike and occupancy, and a smoothed rate map was constructed by dividing the spike map by the occupancy map. Independent rate maps were constructed for the different running directions in the mazes. Only periods in which the animal velocity was >4 cm/s were included. A place field was defined as a continuous region of at least 15 cm^2, where the mean firing rate was >10% of the peak rate in the maze, and the peak firing rate was >3Hz.

Spatial information (SI) for individual place cells was obtained from the linearized rate maps (*97*)

$$SI = \sum_{i=1}^{N} p_i \frac{\lambda_i}{\lambda} \log_2 \frac{\lambda_i}{\lambda}$$

where N is the total number of spatial bins, p_i is the probability of occupancy, and λ_i is the firing rate in the ith spatial bin, and λ the average firing rate of the cell. To measure the experience-dependent change of SI across laps in each condition, the mean and standard deviation of SI across all cells within a condition group were calculated, and the SI was then z-scored for that condition (Fig. 3B).

To identify cells that discriminate between the two sides of the CPP task, we constructed a vector for each cell containing the mean firing rate for every movement interval that the animal spent in each of the two halves. We then performed a one-tailed Wilcoxon rank sum test between the firing vectors in the two halves, and cells for which this comparison was significant were identified as cells with firing specific to that respective half. To quantify discrimination in post-task sleep, pairwise post-task sleep correlations (C_{POST}, see below) were compared between pairs of cells specific for the same half versus pairs of cells that were specific for opposite halves of the environment.

Population vector analysis

To estimate the stability of the population code in the Stim OFF and Stim ON conditions, we performed population vector analysis as described previously (*98*). Briefly, we measured the correlation between the population firing curves in even and in odd trials for each running direction, and we compared this value with the value obtained in surrogate data, in which cell identity was shuffled across even and odd trials.

In addition, we performed a cross-correlation between the population firing curves in even and odd trials separately for each pair of spatial bins (fig. S3, A and B). To estimate representation stability, we computed the "error" defined as the distance between the peak correlation for each spatial bin (each of the columns of the matrix) to the diagonal. To statistically test the directionality of the hippocampal code, we compared the representation stability within a given direction (even versus odd trajectories in the same direction) to the stability of the representation between the two directions of movement (even trajectories in one direction versus odd trajectories in the other direction).

COM shift

To measure PF shifting across laps (Fig. 3D), we used place cells that had PFs on the linear track, and calculated the center-of-mass (COM) of PFs for each lap n (COM_n) (*99*)

$$\mathrm{COM}_n = \frac{\sum_i FR_i x_i}{\sum_i FR_i}$$

where FR_i is the firing rate in the spatial bin i, and x_i is the distance of the spatial bin i from the start of the running trajectory. For each lap, COM shift was then measured as the difference between the first lap and the current lap. To prevent any edge effect, only PFs with a peak location outside the two ends of the track (i.e., defined by 20% of the track length) were included. For place cells with multiple PFs, only the primary PF (with the maximal peak rate) was included.

Place field eccentricity

To calculate the eccentricity of 2D place fields in the CPP task (fig. S12), we first defined the PF boundary on the 2D rate maps. PF was first defined as the area with firing rates larger than 40% of the peak rate. This area was further refined with a series of morphological operations: opening, closing, and infill, with Matlab functions *imopen*, *imclose*, and *imfill*, respectively. PF boundary was then detected on the resulting binarized image using the Matlab function *bwboundaries*. The eccentricity of the detected PF was measured using the Matlab function *regionprops*. For place cells with multiple PFs, only the primary PF (with the maximal peak rate located within the PF boundary with the maximal size) was included. The boundary and nonboundary cells were defined as the cells with the peak of the primary PF located within and outside 15 cm of the wall, respectively.

Phase precession

Phase precession analysis was performed as previously described (*34*, *100*). Circular-linear regression between relative position within the place field and theta phase was applied to calculate the phase-precession slope and correlation strength (*101*). The slope and correlation strength (r^2) of phase precession were derived from this circular-linear regression analysis.

Theta compression

Theta compression analysis was performed as described previously (*102*), independently for the Stim OFF and the Stim ON direction. Briefly, for each pair of overlapping place cells, we computed (i) the distance between their place field peaks and (ii) the theta time lag. To qualify pairs with a significant theta time lag, we computed a coarse cross-correlogram for each cell pair using for durations ±1 s using 5-ms bins. Cell pairs with a peak in this coarse correlogram within ±100 ms with at least five spikes in the peak bin were deemed significant, and other cell pairs were excluded from this analysis. To compute the theta time lag for the selected cell pairs, cross-correlograms were restricted to in-field spikes during running periods of the respective direction (Stim OFF versus Stim ON) using 1-ms bins, and each cross-correlogram was filtered with a bandpass filter between 2 and 30 Hz. The time of the cross-correlogram peak was defined as the theta time lag of the pair of place cells. Theta compression slopes were computed by performing linear regression between the distance between the place field peaks and the theta time lags.

Linearizing positions on the cheeseboard maze

Positions on the cheeseboard maze were linearized along an optimal trajectory, connecting the start box with the three rewarded locations for that session (with the closest reward locations connecting directly to the start box). The animal's linearized position was defined as the relative position along the optimal trajectory closest to the animal's current position. To decode animal position from neuronal activity (for theta sequence and replay analyses), we only retained periods when this linearized position described the current animal's position well, defined as periods in which (i) the error distance between the optimal trajectory and the actual animal position was within 20 cm and (ii) the animal running speed along the optimal trajectory was at least 10 cm/s. To decode replay of familiar routes to previously learned trajectories (fig. S7), we modified the above procedure to define the linearized trajectory: Instead of using the three currently rewarded locations, we used the three rewarded locations from the previous day's session.

Decoding animal position from neuronal activity

To decode position from place cell activity during both theta states and replay, we used a Bayesian reconstruction approach as described previously (*65*). Briefly, we computed the average firing rate probability $\lambda_i^{OFF}(x)$ and $\lambda_i^{ON}(x)$ for each pyramidal cell i at position x as the normalized firing rate curve (spatial bin size: 1.5 cm) of the cell during running epochs in the OFF and the ON direction, respectively,

which constituted the training step in our decoder. To then decode the animal position from the neuronal activity as expressed by the firing rate vector n in a particular window of width τs, we estimate the probability $P(x|n)$

$$P(x|n) = \frac{P(n|x) \cdot P(x)}{P(n)}$$

where spikes are assumed to fire as independent Poisson processes

$$P(n|x) = \prod_{i=1}^{N} \frac{(\lambda_i(x) \cdot \tau)^{n_i}}{n_i!} e^{-\lambda_i(x) \cdot \tau}$$

We applied the same procedure to decode linearized positions on the cheeseboard maze, using periods of movement along the linearized trajectory in which the animal stayed within 20 cm of the optimal trajectory to train the decoder.

Theta sequences

Theta reconstruction matrices

To decode the animal position during theta cycles, we trained Bayesian decoders for the OFF and the ON directions as described above and decoded the animal position during running periods in the OFF and the ON direction, respectively, using τ = 20 ms bins with 5-ms sliding window. We then computed theta reconstruction matrices for each theta cycle while the animal was running in the OFF and the ON direction by dividing the cycle in 500 temporal bins and interpolating the respective (OFF or ON) decoded position. The same procedure was applied to construct theta reconstruction matrices for Stim OFF and Stim ON cheeseboard sessions using the linearized cheeseboard position during periods of movement along the linearized trajectory and stayed within 20 cm of the optimal trajectory.

Theta cycles

To detect theta cycles, we first detected deep and superficial CA1 sublayers using the depth profile of SWRs as described previously (*51*) and selected the deep CA1 channel with the highest theta power in the normalized spectrum. Theta cycles were defined as the peaks in the LFP filtered in the theta band during running periods, and cycles shorter than 100 ms or longer than 200 ms were excluded from further analyses. To correct for the phase shift that can occur in different sublayer depth of the theta channel, the final theta cycles were shifted to ensure that the theta peaks were defined as the phase of maximal uncertainty of the theta reconstruction matrices, where uncertainty u was defined as $u = -\max_x\{P(x|n)\}$. Theta peaks correspond 0° (0 rad.) and 360° (2π rad.) and troughs at 180° (π rad.) and 540° (3π rad.) of theta waves.

Theta sequences

To quantify theta sequence, we applied the measures of quadrant scores and weighted correlation on the theta reconstruction matrices for each cycle. To compute the quadrant score, we divided the central region in the theta reconstruction matrix that was within 50 cm of the animal's current location and within $\frac{1}{2}\pi$ of the theta trough (π) into four equal quadrants, and computed the ratio between the probability within quadrants II and IV (representing directions consistent with the animal's running direction) and the probability within quadrants I and III (representing directions incongruent with the animal's running direction). The weighted correlation measured the correlation between time t and location l weighted by the decoded probability values p within the reconstruction matrix

$$w = \frac{\mathrm{cov}(t, l|p)}{\sqrt{\mathrm{cov}(t, t|p)\mathrm{cov}(l, l|p)}}$$

where

$$\mathrm{cov}(t, l|p) = \frac{\sum_i p_i (t_i - m(t|p))(l_i - m(l|p))}{\sum_i p_i},$$

$$m(t|p) = \frac{\sum_i p_i t_i}{\sum_i p_i}, \text{ and } m(l|p) = \frac{\sum_i p_i l_i}{\sum_i p_i}.$$

The look-ahead index of theta sequences is calculated as in previous reports (*46*). In brief, we recomputed the reconstruction matrix quadrants, this time extending the reach to use all decoded probabilities within 1 m of the animal's current location. The look-ahead index was measured by comparing the probabilities in the quadrants ahead (quadrant IV, future) and behind (quadrant I, past) in the second half of theta phases, as (IV − I)/(IV + I).

Replay analysis

Candidate replay events

We first detected periods of elevated pyramidal cell population activity as defining periods when the instantaneous firing rate (1-ms bins, smoothed with a Gaussian kernel of width 10 ms) was greater than the mean and reached a peak of >2 SD above the mean during the session. Candidate events were between 100 and 500 ms in duration, and they each overlapped with an independently detected SWR; otherwise they were excluded from further analyses.

Quantifying replay

We decoded the animal position for the OFF and the ON direction during each candidate replay event using non-overlapping τ = 20 ms bins (but see fig. S7A for replay results with τ = 40 ms and τ = 60 ms). OFF and ON direction replay were scored independently. To assess replay quality, we computed trajectory scores of the reconstruction matrices (*103*). Briefly, each reconstruction matrix was fitted with a line (of slope a and intercept b) maximizing the average likelihood R that the animal is located within a distance d (set to 22.5 cm) of the linear trajectory defined by the slope a and intercept b

$$R(v, \rho) = \frac{1}{m} \sum_{k=0}^{m-1} P(|pos - (\rho - v \cdot k \cdot \Delta\xi| \leq d))$$

The score of each event was normalized by subtracting the mean and dividing by the standard deviation of a distribution of 500 shuffled scores obtained by independently shifting each of the columns of the event's reconstruction matrix along the spatial dimension by a random distance. An event with a score exceeding the scores of 95% of its shuffled scores was considered significant. To quantify replay, we normalized the proportion of significant events during awake behavior or post-task sleep by subtracting the mean proportion of significant events in pre-task sleep.

In a control analysis, we corrected for the differences in decoding quality between Stim OFF and Stim ON trajectories (fig. S7C). For this purpose, downsampled units before decoding the Stim OFF trajectories by progressively dropping units until the average decoded error in Stim OFF trajectories was larger than the average decoded error in Stim ON trajectories, and we recomputed replay events for Stim OFF trajectories using this more limited sample (fig. S7C).

We performed an additional quantification of replay, where for each replay event we computed the weighted correlation and the jump distance. The weighted correlation measures the correlation coefficient between time and decoded space, and jump distance was defined as the maximum distance between the peak decoded positions in two successive time bins of the same event. Trajectory events were defined as the events with high weighted correlation (>0.6) and low jump distance (<75 cm), and to assess whether the number of trajectory events exceeded chance, we compared the number trajectory events in the recorded data with the number of trajectory events in surrogate data where cell identities were shuffled 500 times. This comparison yielded a P value (the proportion of shuffled datasets with as many or more trajectory events than those observed in the original dataset), and we generated a significance matrix of P values for progressively stricter weighted correlation thresholds (0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95) and jump distance thresholds (75 cm, 62.5 cm, 50 cm, 37.5 cm, 25 cm, 12.5 cm).

Pairwise bias correlations

The above replay analyses test for the replay of stereotypical trajectories and are not directly applicable to the trajectory-free behavior in the

CPP task. To assess whether the order of unit spiking in behavior is preserved in these sessions, we computed the similarity between behavioral sequences and SWR sequences using a pairwise bias correlation method (*65*). Briefly, we computed a bias matrix B_k for each sequence k

$$B_k = \frac{n_k(i,j)}{n_k(i) \cdot n_k(j)}$$

where $n_k(i)$ is the number of spikes emitted by neuron i in the kth sequence, and $n_k(i,j)$ is the number of times neuron i spiked before neuron j in the kth sequence (see fig. S12F for an example). $B_k(i,j)$ therefore reflects the bias of i to spike before j in the kth sequence, taking values between 0 (i never precedes j) and 1 (i always precedes j), with 0.5 representing no bias (i precedes j half of the time). The correlation between two sequences can be estimated as the cosine between their skew-bias matrices as each bias matrix is unwrapped into a vector $2B_k - 2$ (this normalization is to achieve a range of values between −1 and 1, assuring that two mirror sequences would have opposite signs with vectors pointing in opposite directions). The activity of cells that are not common to the two sequences is ignored.

To test whether the order of firing in behavior was preserved in post-task sleep, we computed the mean bias matrix of all sequences taking place in theta cycles during the task in a given recording session. We computed the correlation (cosine) between this bias matrix and each of the bias matrices describing sequences in post-task sleep SWRs. We compared this correlation with a shuffled distribution obtained by recomputing the correlation after shuffling the spikes sequence n = 100 times and deemed as significant those sequences with correlations whose absolute value exceeded 5% of the shuffled distribution. We normalized the proportion of significant events by subtracting the mean proportion of significant events in pre-task sleep.

Coactivity and reactivation analysis

Cofiring

Each spike train was binned in individual theta cycles, and the cofiring of a pair of cells was defined as the Pearson correlation coefficient of the binned firing rates of the two cells across all theta cycles in the recorded session. The matrix of Pearson correlation coefficients was called the cofiring matrix C.

Explained variance

Explained variance (EV) was quantified as previously described (*104*). Briefly, for each session, we computed the proportion of the variance in the population spiking activity in post-task sleep that is explained by task-related activity, after taking into account correlations preexisting in sleep before the experience

$$EV = \left(\frac{r_{C,C_{POST}} - r_{C,C_{PRE}} r_{C_{POST},C_{PRE}}}{\sqrt{\left(1 - r^2_{C,C_{PRE}}\right)\left(1 - r^2_{C_{POST},C_{PRE}}\right)}} \right)^2$$

where C, C_{PRE}, and C_{POST} refer to the cofiring matrix and the correlation matrices across SWRs in pre- and post-task sleep, respectively, and $r_{A,B}$ describes the correlation coefficient between the pairwise correlation matrices A and B. EV was computed for each recorded session separately, and it was compared with the control value of reverse explained variance (REV) obtained by recomputing EV after swapping C_{PRE} and C_{POST}.

Pairwise reactivation

For the pairwise reactivation analyses, we used linear regression between C_{POST} and C_{PRE} to remove the portion of the variance in C_{POST} that can be explained by preexisting patterns, and we correlated the resultant residuals $C_{RESIDUAL}$ with C (*105*).

For the CPP reactivation analysis, we identified cells that had a preference for each of the two sides of the maze by computing each cell's firing rate for every movement period that the animal spent moving in each of the two sides of the maze. Cells with a significant preference (Wilcoxon rank sum test, alpha α = 0.05) were divided according to the sign of their preference between rewarded-side responsive cells and unrewarded-side responsive cells. We then compared $C_{RESIDUAL}$ (see above) in pairs of rewarded-side responsive cells versus pairs of unrewarded-side responsive cells.

Cell assembly analyses

For detecting cell assembly patterns, we used an unsupervised statistical framework based on a hybrid principal components analysis (PCA) followed by independent component analysis (ICA) as previously reported (*49*, *106*–*108*). In brief, spike trains of each neuron were binned in 25-ms intervals for the whole session (including only task periods during movement), the matrix of firing correlation coefficients for all pairs of neurons were constructed. Next, we calculated the number of assemblies on the basis of those principal components whose eigenvalues exceeded the threshold for random firing correlations (using the Marchenko-Pastur law). This method provides a number of significant patterns smaller than the number of neurons. Each of these patterns explained more variance of the spike train correlation matrix than other patterns that would result from independently firing neurons. Independent component analysis (fast-ICA algorithm) was then used to determine for each assembly (component) the vector of weights with which each neuron's firing contributes to that assembly. The strength of each assembly i's activation for a given time bin k was computed as follows

$$S_{ik} = Z_k^T P Z_k$$

where Z is the population activity matrix of the z-scored firing rate of each unit, and P is the outer product of the component i's weights, in which the diagonal has been set to zero so that isolated spikes from individual units do not contribute to S.

Most of the detected assembly patterns consisted of a few neurons with high weights and a large group of neurons with weights around zero. Assembly members were thus considered those cells whose weight exceeded the mean weight of the assembly by two standard deviations (*107*).

To produce the assembly activity profiles in fig. S9B, we took the assembly activation strength during the last hour of cumulative baseline sleep and the first hour of cumulative post-task sleep, excluding periods outside of non-REM sleep. During the task, we took the assembly activation strength during running in Stim OFF and Stim ON trajectories separately. Because the exploration time is variable between sessions, instead of absolute elapsed time, we divided the Stim OFF and Stim ON exploration time into quantiles.

Assembly activation events were detected as peaks in the assembly activation strength exceeding a value of 5 (*109*). The point process of assembly activation events during behavior in the task was used to compute the spatial information content of assembly activity in fig. S4G. The order of assembly activation events in SWRs was analyzed in fig. S9E as follows. For each SWR in which at least three assemblies were active, we correlated the timing of assembly activations within an event to the respective positions (that is, the location of peak activity for a given assembly). We took the absolute value of the Spearman's rank-order correlation coefficient ρ to account for reverse replay. To correct for coefficient variation with the number of assemblies (e.g., fewer points yield higher correlation coefficients) and the day of the recording, we subtracted from each coefficient ρ a baseline value $\rho_0{}^n$ defined as the mean ρ of SWRs recorded pre-task sleep of the same session with a matching number n of assembly activations.

Spectral analysis, cross-frequency coupling, and spike-LFP coupling

To obtain the phase of the theta rhythm, one LFP channel was selected as explained above. LFP was bandpass-filtered in the range of 5 to 15 Hz. Theta phase was then computed using the Hilbert transform of the filtered LFP.

Cross-frequency coupling

To perform spectral analysis at a high resolution in time and frequency, the complex wavelet transform (CWT) of the LFP (or ICs) was calculated using complex Morlet wavelets. Wavelets were calculated using a logarithmically spaced frequency vector in the band of interest (25 to 200 Hz). Phase-amplitude cross-frequency coupling for a given LFP recording was assessed using the modulation index measure (MI) (*110*). Phase time-series were binned into phase intervals and the mean wavelet amplitude was calculated for each of them. The MI was obtained by measuring the divergence of the observed amplitude distribution from the uniform distribution. The statistical significance of the MI values (P value) was assessed by a surrogate analysis (n = 1000 surrogates) with random shifts between the phase and amplitude time series. For the presented plots, grand averages were calculated as the mean across all animals, unless otherwise indicated, and MI reported were significant [$P < 0.01$ compared with surrogate distribution (*34*, *55*)].

Spike-LFP coupling

The phase-locking of spikes to LFP features at each frequency was measured for individual units using the wavelet phase from 25 to 200 Hz (30 logarithmically spaced wavelet scales) at the time of each spike (*34*, *55*). Only neurons that fired at least 100 spikes during the selected task intervals were included in the analysis. Reference LFP was taken 200 to 400 μm away from the electrode, where the unit was recorded to minimize spike energy leakage into the LFP. Modulation indices were calculated using the mean resultant length of the phases, and significance was estimated using the Rayleigh test for nonuniformity ($P < 0.05$) using circular statistics. Preferred frequency of modulation was determined as the largest mean vector length of each significantly modulated neuron. The mean angle of the phases for a given neuron's spikes was taken as the preferred phase.

Independent component and current source density analysis of LFPs

To separate the different sources that contribute to the LFP mixed signal, we used an approach based on independent component analysis (ICA) that has been described and validated previously for hippocampal recordings (*41*, *55*, *111*). Here, we applied ICA to spatially contiguous LFP channels after filtering in the gamma band (25 to 200 Hz). The ICA algorithm (runinca) (*112*) takes a time series of data with dimension equal to the number of recording sites and returns a time series of the same dimensionality, but rotated so that each dimension represents a different IC. The inverse of the mixing matrix that transforms the LFP data into the ICs gives the channel weight of each component that is captured for each recording site. When projected back to the anatomical location of the recording site, this corresponds to the spatial voltage loadings of each IC (*41*, *55*, *111*). Once ICs have been extracted from the raw LFP traces, they can be analyzed as if they were active independently from activities at other locations.

Hierarchical bootstrap

Hierarchical bootstrap (*113*) was performed to analyze data with hierarchical structure. Briefly, bootstrap datasets were created by resampling with replacement following levels of hierarchical order (in the order of animals followed by sessions). The mean of each resampled bootstrap data was calculated each time, in a total of 1000 resampling times. The final statistics were done on the populations of resampled data from the different experimental conditions. P value indicated the test result that the values of one condition were higher than the values of the other condition after controlling for the other nesting variables.

Statistical analyses

Statistical analyses were performed with MATLAB functions or custom-made scripts. No specific analysis was used to estimate minimal population sample or group size, but the number of animals, sessions, and recorded cells was larger or similar to those used in previous related works (*41*, *65*, *66*, *75*, *92*). The unit of analysis was typically identified as single neurons or assemblies. In a few cases, the unit of analysis was sessions or animals, and this is stated in the text. Unless otherwise noted, nonparametric two-tailed Wilcoxon rank sum (equivalent to Mann-Whitney U test) or Wilcoxon signed-rank test was used for unpaired and paired data, respectively. For multiple comparisons following analysis of variance (ANOVA), Tukey's honest significant difference post-hoc test was used. All statistical tests were two-sided, unless otherwise specified. Throughout the manuscript, statistical significance is indicated by asterisks: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

On box plots, the central mark indicates the median; bottom and top edges of the box indicate the 25th and 75th percentiles, respectively; and whiskers extend to the most extreme data points not considered outliers. Outliers are not displayed in some plots but were included in statistical analysis. Owing to experimental design constraints, the experimenter was not blind to the manipulation performed during the experiment (i.e., optogenetic manipulation).

REFERENCES AND NOTES

1. K. Nakazawa *et al*., Requirement for hippocampal CA3 NMDA receptors in associative memory recall. *Science* **297**, 211–218 (2002). doi: 10.1126/science.1071795; pmid: 12040087
2. P. Tovote, J. P. Fadok, A. Lüthi, Neuronal circuits for fear and anxiety. *Nat. Rev. Neurosci.* **16**, 317–331 (2015). doi: 10.1038/nrn3945; pmid: 25991441
3. G. R. I. Barker, E. C. Warburton, When is the hippocampus involved in recognition memory? *J. Neurosci.* **31**, 10721–10731 (2011). doi: 10.1523/JNEUROSCI.6413-10.2011; pmid: 21775615
4. A. V. Beylin *et al*., The role of the hippocampus in trace conditioning: Temporal discontinuity or task difficulty? *Neurobiol. Learn. Mem.* **76**, 447–461 (2001). doi: 10.1006/nlme.2001.4039; pmid: 11726247
5. C. J. MacDonald, K. Q. Lepage, U. T. Eden, H. Eichenbaum, Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron* **71**, 737–749 (2011). doi: 10.1016/j.neuron.2011.07.012; pmid: 21867888
6. E. C. Tolman, Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208 (1948). doi: 10.1037/h0061626; pmid: 18870876
7. E. A. Maguire, K. Woollett, H. J. Spiers, London taxi drivers and bus drivers: A structural MRI and neuropsychological analysis. *Hippocampus* **16**, 1091–1101 (2006). doi: 10.1002/hipo.20233; pmid: 17024677
8. D. Dupret, J. O'Neill, B. Pleydell-Bouverie, J. Csicsvari, The reorganization and reactivation of hippocampal maps predict spatial memory performance. *Nat. Neurosci.* **13**, 995–1002 (2010). doi: 10.1038/nn.2599; pmid: 20639874
9. H. C. Barron *et al*., Neuronal computation underlying inferential reasoning in humans and mice. *Cell* **183**, 228–243.e21 (2020). doi: 10.1016/j.cell.2020.08.035; pmid: 32946810
10. D. Hassabis, D. Kumaran, S. D. Vann, E. A. Maguire, Patients with hippocampal amnesia cannot imagine new experiences. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 1726–1731 (2007). doi: 10.1073/pnas.0610561104; pmid: 17229836
11. J. Widloski, D. J. Foster, Flexible rerouting of hippocampal replay sequences around changing barriers in the absence of global place field remapping. *Neuron* **110**, 1547–1558.e8 (2022). doi: 10.1016/j.neuron.2022.02.002; pmid: 35180390
12. J. O'Keefe, J. Dostrovsky, The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**, 171–175 (1971). doi: 10.1016/0006-8993(71)90358-1; pmid: 5124915
13. H. Eichenbaum, P. Dudchenko, E. Wood, M. Shapiro, H. Tanila, The hippocampus, memory, and place cells: Is it spatial memory or a memory space? *Neuron* **23**, 209–226 (1999). doi: 10.1016/S0896-6273(00)80773-4; pmid: 10399928
14. F. P. Battaglia, K. Benchenane, A. Sirota, C. M. A. Pennartz, S. I. Wiener, The hippocampus: Hub of brain network communication for memory. *Trends Cogn. Sci.* **15**, 310–318 (2011). pmid: 21696996
15. S. A. Josselyn, S. Köhler, P. W. Frankland, Finding the engram. *Nat. Rev. Neurosci.* **16**, 521–534 (2015). doi: 10.1038/nrn4000; pmid: 26289572
16. R. L. Buckner, The role of the hippocampus in prediction and imagination. *Annu. Rev. Psychol.* **61**, 27–48 (2010). doi: 10.1146/annurev.psych.60.110707.163508; pmid: 19958178
17. K. L. Stachenfeld, M. M. Botvinick, S. J. Gershman, The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017). doi: 10.1038/nn.4650; pmid: 28967910
18. H. Eichenbaum, N. J. Cohen, Can we reconcile the declarative memory and spatial navigation views on hippocampal function? *Neuron* **83**, 764–770 (2014). doi: 10.1016/j.neuron.2014.07.032; pmid: 25144874
19. A. M. Wikenheiser, A. D. Redish, Decoding the cognitive map: Ensemble hippocampal sequences and decision making. *Curr. Opin. Neurobiol.* **32**, 8–15 (2015). doi: 10.1016/j.conb.2014.10.002; pmid: 25463559
20. D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory* (Wiley, 1949).
21. K. D. Harris, J. Csicsvari, H. Hirase, G. Dragoi, G. Buzsáki, Organization of cell assemblies in the hippocampus. *Nature* **424**, 552–556 (2003). doi: 10.1038/nature01834; pmid: 12891358
22. M. El-Gaby *et al*., An emergent neural coactivity code for dynamic memory. *Nat. Neurosci.* **24**, 694–704 (2021). doi: 10.1038/s41593-021-00820-w; pmid: 33782620
23. W. E. Skaggs, B. L. McNaughton, M. A. Wilson, C. A. Barnes, Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus* **6**, 149–172 (1996). doi: 10.1002/(SICI)1098-1063(1996)6:2<149::AID-HIPO6>3.0.CO;2-K; pmid: 8797016
24. G. Dragoi, G. Buzsáki, Temporal encoding of place sequences by hippocampal cell assemblies. *Neuron* **50**, 145–157 (2006). doi: 10.1016/j.neuron.2006.02.023; pmid: 16600862

25. D. J. Foster, M. A. Wilson, Hippocampal theta sequences. *Hippocampus* **17**, 1093–1099 (2007). doi: 10.1002/hipo.20345; pmid: 17663452
26. W. E. Skaggs, B. L. McNaughton, Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science* **271**, 1870–1873 (1996). doi: 10.1126/science.271.5257.1870; pmid: 8596957
27. A. K. Lee, M. A. Wilson, Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron* **36**, 1183–1194 (2002). doi: 10.1016/S0896-6273(02)01096-6; pmid: 12495631
28. G. Girardeau, K. Benchenane, S. I. Wiener, G. Buzsáki, M. B. Zugaro, Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci.* **12**, 1222–1223 (2009). doi: 10.1038/nn.2384; pmid: 19749750
29. S. P. Jadhav, C. Kemere, P. W. German, L. M. Frank, Awake hippocampal sharp-wave ripples support spatial memory. *Science* **336**, 1454–1458 (2012). doi: 10.1126/science.1217230; pmid: 22555434
30. I. Gridchyn, P. Schoenenberger, J. O'Neill, J. Csicsvari, Assembly-specific disruption of hippocampal replay leads to selective memory deficit. *Neuron* **106**, 291–300.e6 (2020). doi: 10.1016/j.neuron.2020.01.021; pmid: 32070475
31. Y. Wang, S. Romani, B. Lustig, A. Leonardo, E. Pastalkova, Theta sequences are essential for internally generated hippocampal firing fields. *Nat. Neurosci.* **18**, 282–288 (2015). doi: 10.1038/nn.3904; pmid: 25531571
32. T. Nakashiba, D. L. Buhl, T. J. McHugh, S. Tonegawa, Hippocampal CA3 output is crucial for ripple-associated reactivation and consolidation of memory. *Neuron* **62**, 781–787 (2009). doi: 10.1016/j.neuron.2009.05.013; pmid: 19555647
33. M. I. Schlesiger *et al*., The medial entorhinal cortex is necessary for temporal organization of hippocampal neuronal activity. *Nat. Neurosci.* **18**, 1123–1132 (2015). doi: 10.1038/nn.4056; pmid: 26120964
34. A. Fernández-Ruiz *et al*., Entorhinal-CA3 dual-input control of spike timing in the hippocampus by theta-gamma coupling. *Neuron* **93**, 1213–1226.e5 (2017). doi: 10.1016/j.neuron.2017.02.017; pmid: 28279355
35. B. Lasztóczi, T. Klausberger, Hippocampal place cells couple to three different gamma oscillations during place field traversal. *Neuron* **91**, 34–40 (2016). doi: 10.1016/j.neuron.2016.05.036; pmid: 27387648
36. I. Zutshi, M. Valero, A. Fernández-Ruiz, G. Buzsáki, Extrinsic control and intrinsic computation in the hippocampal CA1 circuit. *Neuron* **110**, 658–673.e5 (2022). doi: 10.1016/j.neuron.2021.11.015; pmid: 34890566
37. J. B. Hales *et al*., Medial entorhinal cortex lesions only partially disrupt hippocampal place cells and hippocampus-dependent place memory. *Cell Rep.* **9**, 893–901 (2014). doi: 10.1016/j.celrep.2014.10.009; pmid: 25437546
38. C. Grienberger, J. C. Magee, Entorhinal cortex directs learning-related changes in CA1 representations. *Nature* **611**, 554–562 (2022). doi: 10.1038/s41586-022-05378-6; pmid: 36323779
39. J. Yamamoto, S. Tonegawa, Direct medial entorhinal cortex input to hippocampal CA1 is crucial for extended quiet awake replay. *Neuron* **96**, 217–227.e4 (2017). doi: 10.1016/j.neuron.2017.09.017; pmid: 28957670
40. J. Dimidschstein *et al*., A viral strategy for targeting and manipulating interneurons across vertebrate species. *Nat. Neurosci.* **19**, 1743–1749 (2016). doi: 10.1038/nn.4430; pmid: 27798629
41. A. Fernández-Ruiz *et al*., Gamma rhythm communication between entorhinal cortex and dentate gyrus neuronal assemblies. *Science* **372**, eabf3119 (2021). doi: 10.1126/science.abf3119; pmid: 33795429
42. J. O'Keefe, M. L. Recce, Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* **3**, 317–330 (1993). doi: 10.1002/hipo.450030307; pmid: 8353611
43. A. M. Wikenheiser, A. D. Redish, Hippocampal theta sequences reflect current goals. *Nat. Neurosci.* **18**, 289–294 (2015). doi: 10.1038/nn.3909; pmid: 25559082
44. A. Johnson, A. D. Redish, Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* **27**, 12176–12189 (2007). doi: 10.1523/JNEUROSCI.3761-07.2007; pmid: 17989284
45. K. Kay *et al*., Constant sub-second cycling between representations of possible futures in the hippocampus. *Cell* **180**, 552–567.e25 (2020). doi: 10.1016/j.cell.2020.01.014; pmid: 32004462
46. W. Tang, J. D. Shin, S. P. Jadhav, Multiple time-scales of decision-making in the hippocampus and prefrontal cortex. *eLife* **10**, e66227 (2021). doi: 10.7554/eLife.66227; pmid: 33683201
47. M. R. Mehta, C. A. Barnes, B. L. McNaughton, Experience-dependent, asymmetric expansion of hippocampal place fields. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 8918–8921 (1997). doi: 10.1073/pnas.94.16.8918; pmid: 9238078
48. L. Genzel *et al*., A consensus statement: Defining terms for reactivation analysis. *Philos. Trans. R. Soc. London Ser. B* **375**, 20200001 (2020). doi: 10.1098/rstb.2020.0001; pmid: 32248790
49. A. Oliva, A. Fernández-Ruiz, F. Leroy, S. A. Siegelbaum, Hippocampal CA2 sharp-wave ripples reactivate and promote social memory. *Nature* **587**, 264–269 (2020). doi: 10.1038/s41586-020-2758-y; pmid: 32968277
50. K. Diba, G. Buzsáki, Forward and reverse hippocampal place-cell sequences during ripples. *Nat. Neurosci.* **10**, 1241–1242 (2007). doi: 10.1038/nn1961; pmid: 17828259
51. R. E. Harvey, H. L. Robinson, C. Liu, A. Oliva, A. Fernandez-Ruiz, Hippocampo-cortical circuits for selective memory encoding, routing, and replay. *Neuron* **111**, 2076–2090.e9 (2023). doi: 10.1016/j.neuron.2023.04.015; pmid: 37196658
52. T. M. George, W. de Cothi, K. L. Stachenfeld, C. Barry, Rapid learning of predictive maps with STDP and theta phase precession. *eLife* **12**, e80663 (2023). doi: 10.7554/eLife.80663; pmid: 36927826
53. A. Alvernhe, E. Save, B. Poucet, Local remapping of place cell firing in the Tolman detour task. *Eur. J. Neurosci.* **33**, 1696–1705 (2011). doi: 10.1111/j.1460-9568.2011.07653.x; pmid: 21395871
54. S. Tanni, W. de Cothi, C. Barry, State transitions in the statistically stable place cell population correspond to rate of perceptual change. *Curr. Biol.* **32**, 3505–3514.e7 (2022). doi: 10.1016/j.cub.2022.06.046; pmid: 35835121
55. E. W. Schomburg *et al*., Theta phase segregation of input-specific gamma patterns in entorhinal-hippocampal networks. *Neuron* **84**, 470–485 (2014). doi: 10.1016/j.neuron.2014.08.051; pmid: 25263753
56. V. Lopes-Dos-Santos *et al*., Parsing hippocampal theta oscillations by nested spectral components during spatial exploration and memory-guided behavior. *Neuron* **100**, 940–952.e7 (2018). doi: 10.1016/j.neuron.2018.09.031; pmid: 30344040
57. L. L. Colgin *et al*., Frequency of gamma oscillations routes flow of information in the hippocampus. *Nature* **462**, 353–357 (2009). doi: 10.1038/nature08573; pmid: 19924214
58. J. Csicsvari, H. Hirase, A. Mamiya, G. Buzsáki, Ensemble patterns of hippocampal CA3-CA1 neurons during sharp wave–associated population events. *Neuron* **28**, 585–594 (2000). doi: 10.1016/S0896-6273(00)00135-5; pmid: 11144366
59. S. J. Middleton, T. J. McHugh, Silencing CA3 disrupts temporal coding in the CA1 ensemble. *Nat. Neurosci.* **19**, 945–951 (2016). doi: 10.1038/nn.4311; pmid: 27239937
60. H. Davoudi, D. J. Foster, Acute silencing of hippocampal CA3 reveals a dominant role in place field responses. *Nat. Neurosci.* **22**, 337–342 (2019). doi: 10.1038/s41593-018-0321-z; pmid: 30664772
61. A. Ecker *et al*., Hippocampal sharp wave-ripples and the associated sequence replay emerge from structured synaptic interactions in a network model of area CA3. *eLife* **11**, e71850 (2022). doi: 10.7554/eLife.71850; pmid: 35040779
62. G. Q. Bi, M. M. Poo, Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *J. Neurosci.* **18**, 10464–10472 (1998). doi: 10.1523/JNEUROSCI.18-24-10464.1998; pmid: 9852584
63. G. M. Wittenberg, S. S.-H. Wang, Malleability of spike-timing-dependent plasticity at the CA3-CA1 synapse. *J. Neurosci.* **26**, 6610–6617 (2006). doi: 10.1523/JNEUROSCI.5388-05.2006; pmid: 16775149
64. R. K. Mishra, S. Kim, S. J. Guzman, P. Jonas, Symmetric spike timing-dependent plasticity at CA3-CA3 synapses optimizes storage and recall in autoassociative networks. *Nat. Commun.* **7**, 11552 (2016). doi: 10.1038/ncomms11552; pmid: 27174042
65. C. Drieu, R. Todorova, M. Zugaro, Nested sequences of hippocampal assemblies during behavior support subsequent sleep replay. *Science* **362**, 675–679 (2018). doi: 10.1126/science.aat2952; pmid: 30409880
66. U. Farooq, G. Dragoi, Emergence of preconfigured and plastic time-compressed sequences in early postnatal development. *Science* **363**, 168–173 (2019). doi: 10.1126/science.aav0502; pmid: 30630930
67. A. Chenani *et al*., Hippocampal CA1 replay becomes less prominent but more rigid without inputs from medial entorhinal cortex. *Nat. Commun.* **10**, 1341 (2019). doi: 10.1038/s41467-019-09280-0; pmid: 30902981
68. D. S. Olton, B. C. Papas, Spatial memory and hippocampal function. *Neuropsychologia* **17**, 669–682 (1979). doi: 10.1016/0028-3932(79)90042-3; pmid: 522981
69. E. R. Wood, P. A. Dudchenko, H. Eichenbaum, The global record of memory in hippocampal neuronal activity. *Nature* **397**, 613–616 (1999). doi: 10.1038/17605; pmid: 10050854
70. J. O'Keefe, L. Nadel, *The Hippocampus as a Cognitive Map* (Oxford Univ. Press, 1978).
71. T. E. J. Behrens *et al*., What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100**, 490–509 (2018). doi: 10.1016/j.neuron.2018.10.002; pmid: 30359611
72. A. S. Gupta, M. A. A. van der Meer, D. S. Touretzky, A. D. Redish, Hippocampal replay is not a simple function of experience. *Neuron* **65**, 695–705 (2010). doi: 10.1016/j.neuron.2010.01.034; pmid: 20223204
73. H. F. Ólafsdóttir, C. Barry, A. B. Saleem, D. Hassabis, H. J. Spiers, Hippocampal place cells construct reward related sequences through unexplored space. *eLife* **4**, e06063 (2015). doi: 10.7554/eLife.06063; pmid: 26112828
74. H. Igata, Y. Ikegaya, T. Sasaki, Prioritized experience replays on a hippocampal predictive map for learning. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2011266118 (2021). doi: 10.1073/pnas.2011266118; pmid: 33443144
75. J. D. Shin, W. Tang, S. P. Jadhav, Dynamics of awake hippocampal-prefrontal replay for spatial learning and memory-guided decision making. *Neuron* **104**, 1110–1125.e7 (2019). doi: 10.1016/j.neuron.2019.09.012; pmid: 31677957
76. D. Bendor, M. A. Wilson, Biasing the content of hippocampal replay during sleep. *Nat. Neurosci.* **15**, 1439–1444 (2012). doi: 10.1038/nn.3203; pmid: 22941111
77. A. A. Carey, Y. Tanaka, M. A. A. van der Meer, Reward revaluation biases hippocampal replay content away from the preferred outcome. *Nat. Neurosci.* **22**, 1450–1459 (2019). doi: 10.1038/s41593-019-0464-6; pmid: 31427771
78. A. K. Gillespie *et al*., Hippocampal replay reflects specific past experiences rather than a plan for subsequent choice. *Neuron* **109**, 3149–3163.e6 (2021). doi: 10.1016/j.neuron.2021.07.029; pmid: 34450026
79. K. C. Bittner, A. D. Milstein, C. Grienberger, S. Romani, J. C. Magee, Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science* **357**, 1033–1036 (2017). doi: 10.1126/science.aan3846; pmid: 28883072
80. L. Roux, B. Hu, R. Eichler, E. Stark, G. Buzsáki, Sharp wave ripples during learning stabilize the hippocampal spatial map. *Nat. Neurosci.* **20**, 845–853 (2017). doi: 10.1038/nn.4543; pmid: 28394323
81. K. Liu, J. Sibille, G. Dragoi, Preconfigured patterns are the primary driver of offline multi-neuronal sequence replay. *Hippocampus* **29**, 275–283 (2019). doi: 10.1002/hipo.23034; pmid: 30260526
82. A. D. Grosmark, G. Buzsáki, Diversity in neural firing dynamics supports both rigid and learned hippocampal sequences. *Science* **351**, 1440–1443 (2016). doi: 10.1126/science.aad1935; pmid: 27013730
83. A. Noguchi, R. Huszár, S. Morikawa, G. Buzsáki, Y. Ikegaya, Inhibition allocates spikes during hippocampal ripples. *Nat. Commun.* **13**, 1280 (2022). doi: 10.1038/s41467-022-28890-9; pmid: 35277500
84. L. Muessig, M. Lasek, I. Varsavsky, F. Cacucci, T. J. Wills, Coordinated emergence of hippocampal replay and theta sequences during post-natal development. *Curr. Biol.* **29**, 834–840.e4 (2019). doi: 10.1016/j.cub.2019.01.005; pmid: 30773370
85. K. A. Bolding, J. Ferbinteanu, S. E. Fox, R. U. Muller, Place cell firing cannot support navigation without intact septal circuits. *Hippocampus* **30**, 175–191 (2020). doi: 10.1002/hipo.23136; pmid: 31301167
86. D. Robbe, G. Buzsáki, Alteration of theta timescale dynamics of hippocampal place cells by a cannabinoid is associated with memory impairment. *J. Neurosci.* **29**, 12597–12605 (2009). doi: 10.1523/JNEUROSCI.2407-09.2009; pmid: 19812334
87. N. T. M. Robinson *et al*., Medial entorhinal cortex selectively supports temporal coding by hippocampal neurons. *Neuron* **94**, 677–688.e6 (2017). doi: 10.1016/j.neuron.2017.04.003; pmid: 28434800
88. N. M. van Strien, N. L. M. Cappaert, M. P. Witter, The anatomy of memory: An interactive overview of the parahippocampal-hippocampal network. *Nat. Rev.*

Neurosci. **10**, 272–282 (2009). doi: 10.1038/nrn2614; pmid: 19300446
89. S. Melzer *et al*., Long-range-projecting GABAergic neurons modulate inhibition in hippocampus and entorhinal cortex. *Science* **335**, 1506–1510 (2012). doi: 10.1126/science.1217139; pmid: 22442486
90. J. Basu *et al*., Gating of hippocampal activity, plasticity, and memory by entorhinal cortex long-range inhibition. *Science* **351**, aaa5694 (2016). doi: 10.1126/science.aaa5694; pmid: 26744409
91. A. Fernández-Ruiz *et al*., Long-duration hippocampal sharp wave ripples improve memory. *Science* **364**, 1082–1086 (2019). doi: 10.1126/science.aax0758; pmid: 31197012
92. A. Oliva, A. Fernández-Ruiz, G. Buzsáki, A. Berényi, Spatial coding and physiological properties of hippocampal neurons in the Cornu Ammonis subregions. *Hippocampus* **26**, 1593–1607 (2016). doi: 10.1002/hipo.22659; pmid: 27650887
93. A. Mathis *et al*., DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018). doi: 10.1038/s41593-018-0209-y; pmid: 30127430
94. M. Stimberg, R. Brette, D. F. Goodman, Brian 2, an intuitive and efficient neural simulator. *eLife* **8**, e47314 (2019). doi: 10.7554/eLife.47314; pmid: 31429824
95. M. Pachitariu, N. A. Steinmetz, S. N. Kadir, M. Carandini, K. D. Harris, Fast and accurate spike sorting of high-channel count probes with KiloSort. *Adv. Neural Inf. Process. Syst.* **29**, 4448–4456 (2016).
96. P. C. Petersen, J. H. Siegle, N. A. Steinmetz, S. Mahallati, G. Buzsáki, CellExplorer: A framework for visualizing and characterizing single neurons. *Neuron* **109**, 3594–3608.e2 (2021). doi: 10.1016/j.neuron.2021.09.002; pmid: 34592168
97. W. Skaggs, B. McNaughton, K. Gothard, An information-theoretic approach to deciphering the hippocampal code. *Adv. Neural Inf. Process. Syst.* **5**, 1030–1037 (1992).
98. S. Leutgeb *et al*., Independent codes for spatial and episodic memory in hippocampal neuronal ensembles. *Science* **309**, 619–623 (2005). doi: 10.1126/science.1114037; pmid: 16040709
99. C. Dong, A. D. Madar, M. E. J. Sheffield, Distinct place cell dynamics in CA1 and CA3 encode experience in new environments. *Nat. Commun.* **12**, 2977 (2021). doi: 10.1038/s41467-021-23260-3; pmid: 34016996
100. F. Sharif, B. Tayebi, G. Buzsáki, S. Royer, A. Fernandez-Ruiz, Subcircuits of deep and superficial CA1 place cells support efficient spatial coding across heterogeneous environments. *Neuron* **109**, 363–376.e6 (2021). doi: 10.1016/j.neuron.2020.10.034; pmid: 33217328
101. R. Kempter, C. Leibold, G. Buzsáki, K. Diba, R. Schmidt, Quantifying circular-linear associations: Hippocampal phase precession. *J. Neurosci. Methods* **207**, 113–124 (2012). doi: 10.1016/j.jneumeth.2012.03.007; pmid: 22487609
102. K. Diba, G. Buzsáki, Hippocampal network dynamics constrain the time lag between pyramidal cells across modified environments. *J. Neurosci.* **28**, 13448–13456 (2008). doi: 10.1523/JNEUROSCI.3824-08.2008; pmid: 19074018
103. T. J. Davidson, F. Kloosterman, M. A. Wilson, Hippocampal replay of extended experience. *Neuron* **63**, 497–507 (2009). doi: 10.1016/j.neuron.2009.07.027; pmid: 19709631
104. H. S. Kudrimoti, C. A. Barnes, B. L. McNaughton, Reactivation of hippocampal cell assemblies: Effects of behavioral state, experience, and EEG dynamics. *J. Neurosci.* **19**, 4090–4101 (1999). doi: 10.1523/JNEUROSCI.19-10-04090.1999; pmid: 10234037
105. B. Giri, H. Miyawaki, K. Mizuseki, S. Cheng, K. Diba, Hippocampal reactivation extends for several hours following novel experience. *J. Neurosci.* **39**, 866–875 (2019). doi: 10.1523/JNEUROSCI.1950-18.2018; pmid: 30530857
106. V. Lopes-dos-Santos, S. Ribeiro, A. B. L. Tort, Detecting cell assemblies in large neuronal populations. *J. Neurosci. Methods* **220**, 149–166 (2013). doi: 10.1016/j.jneumeth.2013.04.010; pmid: 23639919
107. G. M. van de Ven, S. Trouche, C. G. McNamara, K. Allen, D. Dupret, Hippocampal offline reactivation consolidates recently formed cell assembly patterns during sharp wave-ripples. *Neuron* **92**, 968–974 (2016). doi: 10.1016/j.neuron.2016.10.020; pmid: 27840002
108. R. Todorova, M. Zugaro, Isolated cortical computations during delta waves support memory consolidation. *Science* **366**, 377–381 (2019). doi: 10.1126/science.aay0616; pmid: 31624215
109. V. J. Oberto *et al*., Distributed cell assemblies spanning prefrontal cortex and striatum. *Curr. Biol.* **32**, 1–13.e6 (2022). doi: 10.1016/j.cub.2021.10.007; pmid: 34699783
110. A. B. L. Tort *et al*., Dynamic cross-frequency couplings of local field potential oscillations in rat striatum and hippocampus during performance of a T-maze task. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 20517–20522 (2008). doi: 10.1073/pnas.0810524105; pmid: 19074268
111. N. Benito *et al*., Spatial modules of coherent activity in pathway-specific LFPs in the hippocampus reflect topology and different modes of presynaptic synchronization. *Cereb. Cortex* **24**, 1738–1752 (2014). doi: 10.1093/cercor/bht022; pmid: 23395845
112. A. J. Bell, T. J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* **7**, 1129–1159 (1995). doi: 10.1162/neco.1995.7.6.1129; pmid: 7584893
113. V. Saravanan, G. J. Berman, S. J. Sober, Application of the hierarchical bootstrap to multi-level data in neuroscience. *Neuron Behav. Data Anal. Theory* **3**, 1–25 (2020). pmid: 33644783
114. R. harvey *et al*., ayalab1/neurocode: Release (v1.0.0), version 1.0.0, Zenodo (2023); https://doi.org/10.5281/zenodo.7819979.

ACKNOWLEDGMENTS

The authors thank R. Harvey, H. Robison, L. Sjulson, J. Goldberg, I. Ellwood, W. Sun, M. Zugaro, and D. Nitz for useful comments on the manuscript and M. R. Warden for providing access to the confocal microscope. **Funding:** This work was supported by NIH grant 4R00MH120343, a Sloan Fellowship, a Whitehall Research Grant, and a Klingenstein-Simons Fellowship (A.F.-R.); NIH grant 4R00MH122582 (A.O.); an Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship (a Schmidt Futures program) and a Philippe Foundation grant (R.T.); and a Klarman Fellowship (W.T.). **Author contributions:** All authors conceived of and designed the experiments and analyses and wrote the manuscript. C.L., W.T., and A.F.-R. performed the experiments. R.T., C.L., W.T., and A.F.-R. analyzed the data. R.T. performed the simulations. First authorship order was determined by coin toss. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. Custom scripts used in this study can be downloaded from Zenodo (*114*).
License information: Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adi8237
Figs. S1 to S14
MDAR Reproducibility Checklist

Submitted 19 May 2023; accepted 21 August 2023
10.1126/science.adi8237

RESEARCH ARTICLE

SOLAR CELLS

Stabilized hole-selective layer for high-performance inverted p-i-n perovskite solar cells

Zhen Li[1]†, Xianglang Sun[1, 2]†, Xiaopeng Zheng[3, 4]†, Bo Li[1]†, Danpeng Gao[1], Shoufeng Zhang[1], Xin Wu[1], Shuai Li[1], Jianqiu Gong[1], Joseph M. Luther[3], Zhong'an Li[2]*, Zonglong Zhu[1]*

P-i-n geometry perovskite solar cells (PSCs) offer simplified fabrication, greater amenability to charge extraction layers, and low-temperature processing over n-i-p counterparts. Self-assembled monolayers (SAMs) can enhance the performance of p-i-n PSCs but ultrathin SAMs can be thermally unstable. We report a thermally robust hole-selective layer comprised of nickel oxide (NiO_x) nanoparticle film with a surface-anchored (4-(3,11-dimethoxy-7H-dibenzo[c,g]carbazol-7-yl)butyl)phosphonic acid (MeO-4PADBC) SAM that can improve and stabilize the NiO_x/perovskite interface. The energetic alignment and favorable contact and binding between NiO_x/MeO-4PADBC and perovskite reduced the voltage deficit of PSCs with various perovskite compositions and led to strong interface toughening effects under thermal stress. The resulting 1.53–electron-volt devices achieved 25.6% certified power conversion efficiency and maintained >90% of their initial efficiency after continuously operating at 65 degrees Celsius for 1200 hours under 1-sun illumination.

Phosphonic acid self-assembled monolayers (SAMs) with a carbazole core have enabled performance advances in perovskite solar cells (PSCs), for both single-junction (*1*–*4*) and perovskite-based tandem solar cells (*5*–*9*), because of their high hole selectivity, fast hole transfer rate, and low interfacial trap state density (*10*–*13*). However, compared with conventional polymeric and metal oxide hole transporting materials, SAM-based PSCs have exhibited poorer thermal stability (*14*–*18*). The investigation of the device's operational stability, with electrical bias under maximum power point (MPP) operation or at open-circuit voltage (V_{OC}), under elevated temperatures (65° to 85°C) is critical to improve the confidence in their stability and to meet the qualification of international stability standards (i.e., International Summit on Organic Photovoltaic Stability (ISOS) and International Electrotechnical Commission 61215 standards) (*19*). Whereas most studies on the SAM-based PSCs have reported the operational stability at room temperature or enhanced the device's durability under thermal stress by stabilizing perovskite surface and bulk (*20*, *21*), the degradation effect of SAM-forming molecules under elevated temperature (>65°C) have rarely been discussed (fig. S1 and table S1). The thermal stability of SAM-forming molecules depends largely on their bonding to the selected substrates, as the bond between the anchoring group and the spacer of the molecule can be broken through temperature-induced desorption (*22*–*25*).

We report a new SAM, (4-(3,11-dimethoxy-7H-dibenzo[c,g]carbazol-7-yl)butyl)phosphonic acid (MeO-4PADBC), which we anchored to the NiO_x film to fabricate inverted p-i-n PSCs (fig. S2). The NiO_x/MeO-4PADBC exhibited an optimal dipole moment and amenable surface for favorable contact with perovskite that resulted in an ideal energetic alignment, fast hole extraction, and low defect density. Such an interface configuration also immobilized the SAM molecules at the NiO_x/perovskite interface and produced a robust hole-selective layer (HSL) for thermally stable PSCs that had thermal-degradation activation energy ~3 times greater than that for ITO/MeO-4PADBC. These synergetic effects enabled 1.53-electron volt (eV) p-i-n PSCs V_{OC} of 1.19 V (95% of calculated potential) and a verified power conversion efficiency (PCE) of 25.6%. 1.68- and 1.80-eV–wide bandgap perovskite composition devices also showed encouraging PCEs of 22.7 and 20.1%, respectively. Moreover, 1.53-eV PSCs using NiO_x/MeO-4PADBC HSL maintained >90% of initial efficiency in long-term operational stability tests under 65°C for 1200 hours, and extrapolation of the Arrhenius energy indicates that the solar cells should maintain 80% of initial efficiency for more than 10 months at 25°C.

SAM design and synthesis

Modulating the terminal functional group is an effective way to tune the interfacial interactions between SAMs and perovskite. For example, the introduction of two methoxy gr(OMe) into [2-(9H-carbazol-9-yl)ethyl]phosph acid (2PACz) enabled a new SAM, [2-(3,6-dimethoxy-9H-carbazol-9-yl)ethyl]phosphonic acid (MeO-2PACz), which improved the interfacial contact and enabled a higher efficiency of p-i-n PSC compared with 2PACz (*10*). Nonetheless, the OMe substitution on the carbazole core also caused a decrease in dipole moment from ~2.0 D of 2PACz to ~0.2 D of MeO-2PACz, which in turn led to the offset between the highest occupied molecular orbital (HOMO) of the SAM molecule and the valence band maximum (VBM) of the perovskite (*10*, *26*). This phenomenon can be ascribed to high planarity and symmetry of the carbazole structure; the incorporation of two OMe groups with opposite directions of dipole moment resulted in a net molecular dipole moment of MeO-2PACz close to zero (Fig. 1A). We addressed this issue by using a non-coplanar screw-shaped dibenzo[c,g]carbazole (DBC) unit as the core to reduce the negative effect on the dipole moment when introducing OMe groups, affording a new SAM of MeO-4PADBC (Fig. 1B). Density functional theory (DFT) calculations verified that MeO-4PADBC only has a slightly decreased dipole moment (2.4 D) compared with that of (4-(7H-dibenzo[c,g]carbazol-7-yl)butyl)phosphonic acid (4PADBC) (2.9 D), which is quite different from those obtained for carbazole-based SAMs (figs. S3 and S4).

The synthetic route of MeO-4PADBC is shown in fig. S5, along with structural characterizations presented in figs. S6 to S9. The calculated HOMO and lowest unoccupied molecular orbital distributions are shown in fig. S10, with values of −4.91 and −1.15 eV, respectively. To gain a deeper insight of the SAM molecule structure, we grew a single crystal of 3,11-dimethoxy-7H-dibenzo[c,g]carbazole (MeO-DBC, CCDC number: 2279245), which lacks the anchoring group that would disrupt intermolecular interactions. As shown in fig. S11, A and B, two naphthalene rings are located on different sides of the pyrrole ring and exhibit a dihedral angle of 12.44°, which greatly reduces the planarity and symmetry of the skeleton structure to enable a negligible effect on the dipole moment when introducing OMe groups. Moreover, the MeO-DBC core had a slipped π-stacked packing motif with strong C-H···π (2.72 Å) and π-π (3.79 Å) interactions (fig. S11C), which could induce a highly ordered one-dimensional assembly that favors a dense, tilted, highly ordered monolayer on the substrate (fig. S11D) (*9*, *27*, *28*).

We then calculated the interfacial binding energies between SAMs and perovskite, showing that MeO-4PADBC has a stronger binding with perovskite with a total binding energy (E_b) of −7.19 eV, compared with that of −5.27 eV for MeO-2PACz (fig. S12). The main electrostatic interaction of the calculated SAMs is the O atoms from OMe groups with the Pb from

[1]Department of Chemistry, City University of Hong Kong, Kowloon 999077, Hong Kong. [2]State Key Laboratory of Materials Processing and Die & Mould Technology, Key Laboratory for Material Chemistry of Energy Conversion and Storage, Ministry of Education, School of Chemistry and Chemical Engineering, Huazhong University of Science and Technology, Wuhan, 430074, P.R. China. [3]National Renewable Energy Laboratory, Golden, CO 80401, USA. [4]Center of Materials Science and Optoelectronics Engineering, College of Materials Science and Optoelectronic Technology, University of Chinese Academy of Science, Beijing, 100049, P.R. China.
*Corresponding author. Email: zonglzhu@cityu.edu.hk (Z.Z.); lizha@hust.edu.cn (Zho.L)
†These authors contributed equally to this work.

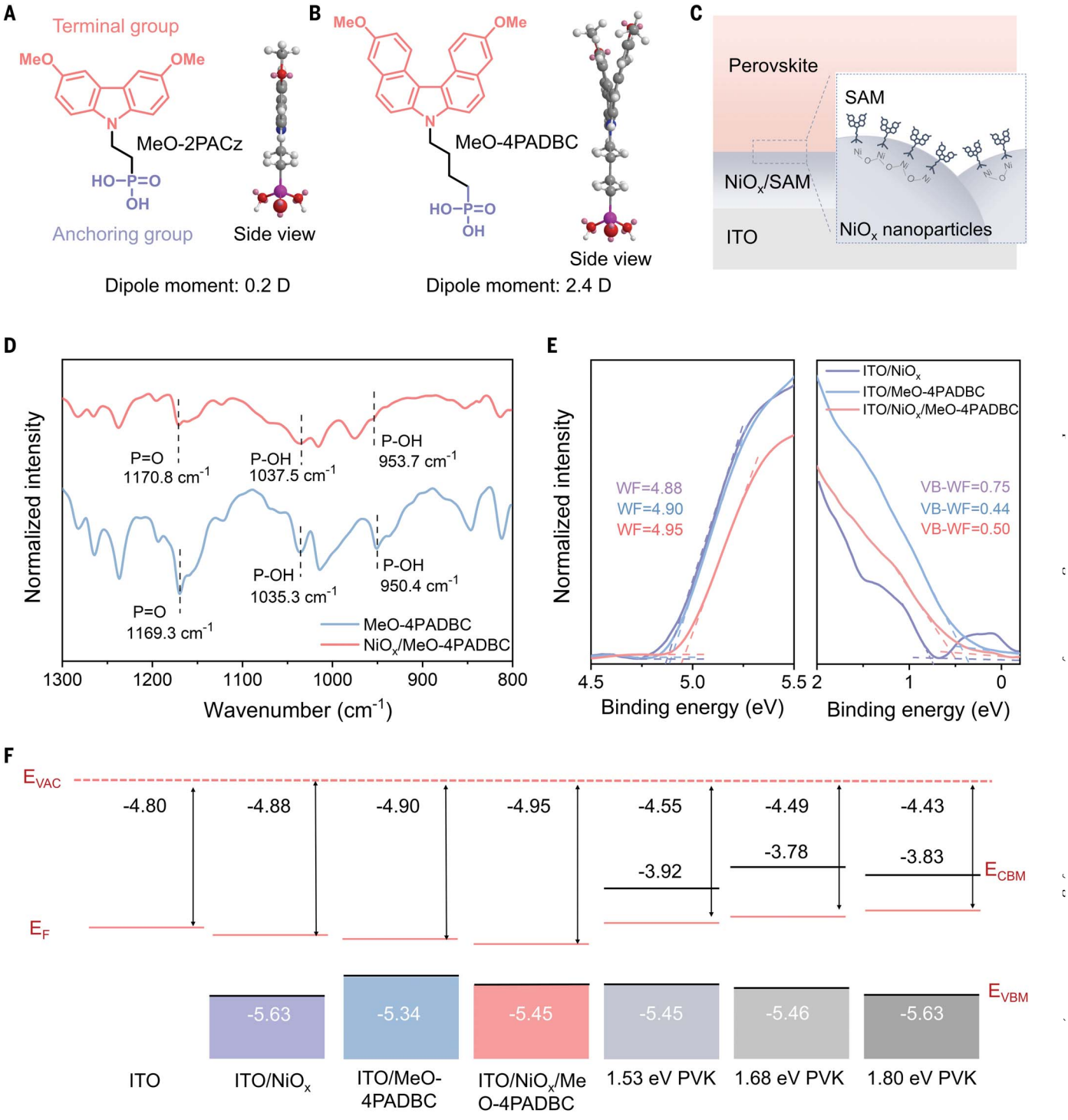


Fig. 1. Molecular structure and electrical properties of HSLs. Molecular structure and side view of (**A**) MeO-2PACz and (**B**) MeO-4PADBC. (**C**) Schematic illustration of MeO-4PADBC anchoring on NiO_x nanoparticle as the HSL in PSC. (**D**) FTIR spectra of MeO-4PADBC and NiO_x/MeO-4PADBC. (**E**) UPS spectra of ITO substrates covered by NiO_x, MeO-4PADBC, and NiO_x/MeO-4PADBC. (Left) UPS spectra around the secondary electron cutoff (WF, work function); (right) UPS spectra in the valence band (VB) region. (**F**) Schematic representation of the band edge positions of the studied HSLs based on values from UPS measurements, referenced to the vacuum level. E_F and E_{VAC} represent Fermi and vacuum levels, respectively. E_{VBM} and E_{CBM} represent the energy of valence band maximum and conduction band minimum, respectively.

perovskite, however MeO-4PADBC has a more energetically favorable contact with perovskite compared with MeO-2PACz, which can be ascribed to the more concentrated electron distribution of the coordinated OMe group on MeO-4PADBC than that on MeO-2PACz (fig. S13). Moreover, Bader charge analysis also suggests a stronger Pb-O interaction between MeO-4PADBC and perovskite with a shorter calculated interaction length of 4.7 Å than that of MeO-2PACz of 4.9 Å (fig. S14), which can be further confirmed by the results

of ^{1}H NMR spectra of the SAMs in DMSO-d_6 with or without mixing with PbI_2 (fig. S15).

Hole-selective layer applications

We compared MeO-4PADBC SAM and NiO_x/MeO-4PADBC as the HSLs for p-i-n PSCs (Fig. 1C). Fourier-transform infrared (FTIR) spectra were then collected by scraping the MeO-4PADBC and NiO_x/MeO-4PADBC films off the substrates to amplify the signal, in which the P=O (1169 cm^{-1}) and P-OH (1035 and 950 cm^{-1}) absorption peaks show a clear shift (Fig. 1D), indicating the formation of a chemical bond (*29*). This can be further confirmed by the changes of the Ni *2p* core level of NiO_x and NiO_x/MeO-4PADBC films deposited on the indium tin oxide (ITO) substrates obtained by x-ray photoelectron spectroscopy (XPS) spectra (fig. S16). The crystal structure and absorbance of representative perovskite film with a composition of $Cs_{0.05}FA_{0.85}MA_{0.1}PbI_3$ on both substrates remained unchanged (figs. S17 and S18). However, the perovskite deposited on ITO/NiO_x/MeO-4PADBC substrate had larger crystal domains than those on control substrates (fig. S19). This is because of the denser anchoring of SAM molecules onto the ITO/NiO_x substrate through more robust tridentate binding absorption, which can not only reduce the surface roughness of ITO/NiO_x substrate (*30*–*32*) (figs. S20 and S21), but also lead to a more hydrophobic surface (fig. S22) and stronger interaction between perovskite and SAM molecules (*33*–*35*). These synergetic effects contribute to facilitating the perovskite crystal nucleation and growth, thus enhancing the perovskite crystallization.

We further used ultraviolet photoelectron spectroscopy (UPS) to assess the energetic alignment of different substrates relative to perovskite absorbers with different bandgaps (1.53, 1.68, and 1.8 eV, see Fig. 1E and fig. S23), and the results are summarized in Fig. 1F. The HOMO energy levels are −5.34 eV for ITO/MeO-4PADBC and −5.45 eV ITO/NiO_x/MeO-4PADBC, respectively, and the work function (Φ) of the latter substrate is deeper than that of the former one (−4.95 versus −4.90 eV), indicative of better energetic alignment with different perovskite absorbers. However, upon application of the carbazole-based SAM molecules onto the ITO/NiO_x surface, the ITO/NiO_x/MeO-2PACz system exhibited an upward shift of the Φ, in contrast to the ITO/NiO_x/2PACz substrate (*26*, *36*) (fig. S24). This phenomenon can be attributed to the incorporation of OMe groups on 2PACz, which considerably reduced the dipole moment. This reduction can be explained by the highly planar carbazole motif, as discussed earlier. Time-resolved photoluminescence (TRPL) decay data showed a decreased carrier lifetime from 925.5 ns for the perovskite on ITO to 42.9 and 32.3 ns for the perovskite on ITO/MeO-4PADBC and ITO/NiO_x/MeO-4PADBC, respectively, indicating a more facilitated hole extraction resulted by the ITO/NiO_x/MeO-4PADBC substrate (fig. S25 and table S2).

Solar cell performance and characterization

We further evaluated the photovoltaic performance of PSCs with a p-i-n device configuration: glass/ITO/HSL/perovskite/two-dimensional (2D) passivation layer/C_{60}/bathocuproine (BCP)/Ag (Fig. 2A). The 2D passivation layer was a mixture of 4-trifluorophenylethylammonium iodide (CF_3-PEAI) and methylammonium iodide (MAI) with an optimal volume ratio of 3:1 (*4*). The cross-section scanning electron microscopy (SEM) images of the PSC (1.53 eV) showed a thickness of ~715 nm for the perovskite films (fig. S26). The device with NiO_x/MeO-4PADBC as the HSL (1.53 eV) showed negligible hysteresis and had a high verified PCE of 25.6% for a mask area of 0.0414 cm^2 (figs. S27 and S28), with a V_{OC} of 1.19 V, a short-circuit current density (J_{SC}) of 25.4 mA cm^{-2}, and a fill factor (FF) of 84.6%, which outperformed the MeO-4PADBC–based device (PCE = 24.2% with V_{OC} = 1.16 V, J_{SC} = 25.4 mA cm^{-2}, and FF = 82.1%). The NiO_x control device without a SAM only showed a PCE of 21.6%, attributed to a mismatch of energetic alignment and high surface defect density on NiO_x film (*37*, *38*) (Fig. 2B). Moreover, we have also conducted HSL engineering by applying carbazole-based SAM anchoring on NiO_x. The V_{OC} of the NiO_x/MeO-2PACz–based device (1.15 V) is lower than that of the device with NiO_x/2PACz (1.17 V) (figs. S29, A and B). This issue was addressed by replacing the carbazole motif with a nonplanar DBC core in SAM molecules (fig. S29, C and D). To better understand the mechanism of the improved V_{OC} and FF with the use of NiO_x/MeO-4PADBC as HSL, we further conducted the FF loss (supplementary text and fig. S30) and V_{OC} loss calculations (supplementary text, fig. S31, and table S3). We observed that the improved V_{OC} and FF of NiO_x/MeO-4PADBC were mainly attributed to the suppressed nonradiative recombination loss, indicating reduced trap state density at the HSL/perovskite interface.

Our NiO_x/MeO-4PADBC strategy also worked effectively for 1.68- and 1.80-eV PSCs, resulting in PCEs of 22.7 and 20.1%, respectively (Fig. 2C, fig. S32, and table S4). Steady-state power output (SPO) confirmed the reliability of three bandgap devices, with stabilized PCEs of 25.5, 22.3, and 19.5% for 1.53-, 1.68- and 1.80-eV, respectively (Fig. 2D). Additionally, the calculated J_{SC} values from external quantum efficiency (EQE) of the champion devices were consistent with those extracted from the current density-voltage (*J-V*) measurements (fig. S33), and the derivatives of EQE spectra can further confirm the perovskite bandgaps applied here (fig. S34). We further presented the EQE with internal quantum efficiency (IQE) results of representative 1.53-eV device with HSLs of NiO_x, MeO-4PADBC, and NiO_x/MeO-4PADBC in fig. S35. The IQE between 550 nm to 700 nm is spectrally flat and approaches nearly 100% for both of MeO-4PADBC and NiO_x/MeO-4PADBC–based devices, indicating efficient charge collection and transfer at the perovskite interface achieved by the incorporation of SAM (*12*, *39*).

The defect density profiles were then studied to identify enhanced photovoltaic performance through the space charge limit current (SCLC) method (fig. S36). The hole-only devices with NiO_x/MeO-4PADBC showed the lowest defect density of 1.96×10^{15} cm^{-3}, compared with those with MeO-4PADBC (2.91×10^{15} cm^{-3}) and NiO_x (3.81×10^{15} cm^{-3}). In addition, the decreased slope of the light intensity dependence V_{OC} plot for NiO_x/MeO-4PADBC-based devices supported a reduced interfacial trap density at the HSL/perovskite interface (*40*) (fig. S37). We note that PSCs based on NiO_x/MeO-4PADBC further demonstrated a lower leakage current than those on other HSLs (fig. S38), which we attributed to the compact NiO_x layer preventing the perovskite from contacting ITO through the pinholes in the ultrathin SAM layer. These results demonstrated that the NiO_x/MeO-4PADBC HSL effectively increased the V_{OC} of the PSCs.

To quantify the interface losses, quasi-Fermi level splitting (QFLS) analysis for partial cell stacks was conducted. A laser wavelength of 375 nm was used to illuminate a PSC with 1-sun equivalent intensity by accommodating the generated current near J_{SC} under a standard solar simulator (*41*, *42*) (supplementary text, fig. S39). As shown in Fig. 2E, the QFLS of ITO/NiO_x/MeO-4PADBC/perovskite stack was comparable to the glass/perovskite stack, with implied V_{OC} of 1.17 V versus 1.18 V for 1.53-eV perovskite film, 1.26 V versus 1.27 V for 1.68-eV perovskite film, and 1.35 V versus 1.36 V for 1.80-eV perovskite film, respectively. These results indicated a low voltage loss on the interface between NiO_x/MeO-4PADBC and perovskites.

We further performed the QFLS measurements on the ITO/HSL/perovskite/passivation layer/electron-transporting layer (ETL) stacks (Fig. 2F), with the V_{OC}s of the PSCs listed as a comparison. The differences between the QFLS of PSCs on ITO/NiO_x/MeO-4PADBC and the V_{OC} extracted from related *J-V* measurements were comparable, demonstrating spatially flat Fermi levels throughout the device and low energy offset on the HSL for carrier extraction (*27*, *28*). It is noteworthy that the V_{OC} of our 1.53-eV devices (1.19 V) reached 95% of their calculated potential and the V_{OC} of our 1.68-eV (1.25 V) and 1.80-eV (1.34 V) devices also approached 90% of the calculated potential (Fig. 2G).

PSC stability studies

Previous reports have demonstrated that SAMs desorb under thermal stress from the anchored

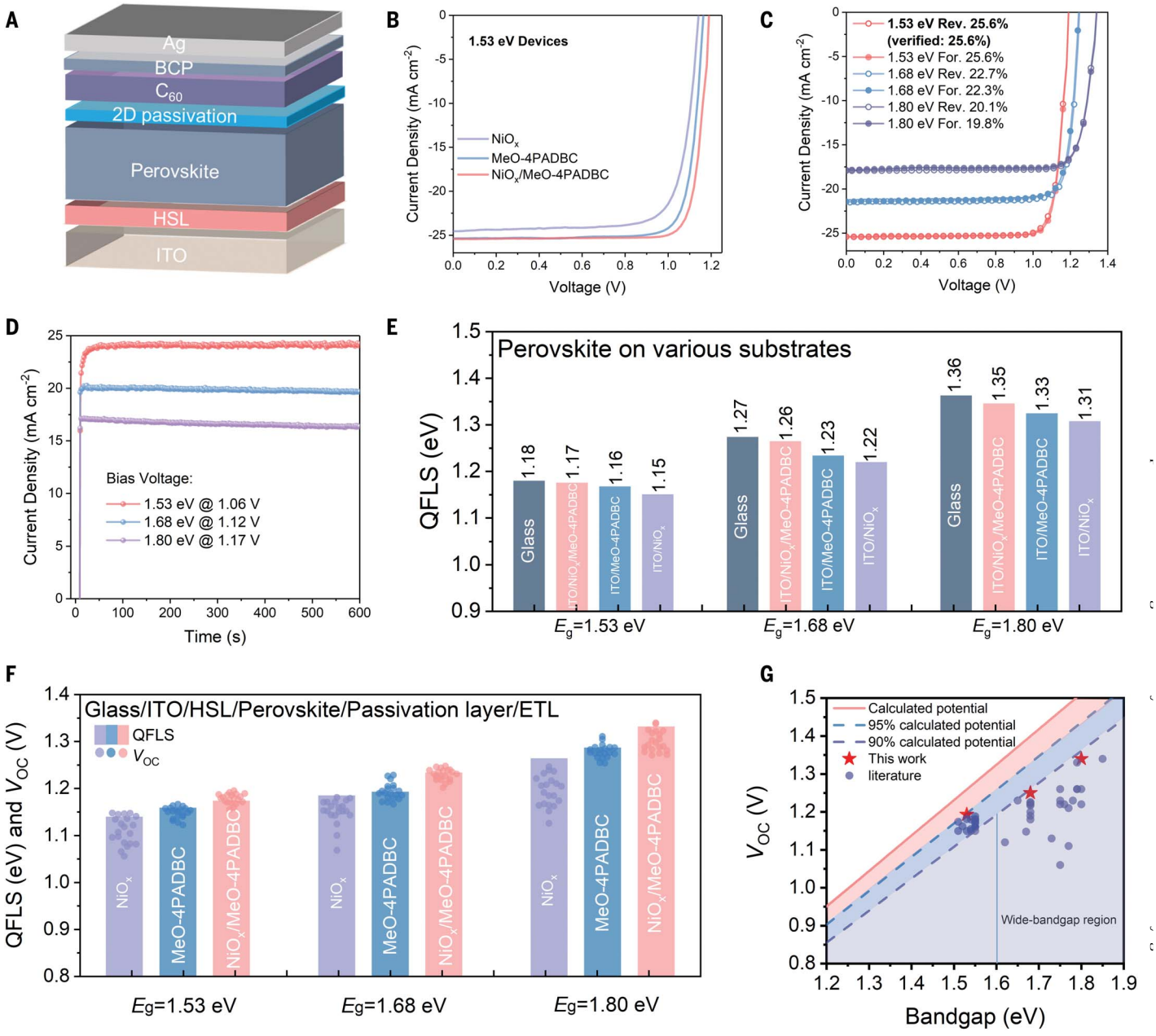


Fig. 2. Photovoltaic performance of PSCs with different HSLs. (**A**) Schematic illustration of p-i-n PSC. (**B**) *J-V* curves of the best-performing 1.53 eV devices with NiO_x, MeO-4PADBC, and NiO_x/MeO-4PADBC as HSL. (**C**) *J-V* curves of the best-performing NiO_x/MeO-4PADBC-based devices with bandgaps of 1.53, 1.68, and 1.80 eV (Rev., reverse scan; For., forward scan). (**D**) SPO at the MPP for the best-performing NiO_x/MeO-4PADBC-based PSCs with three bandgaps of perovskite absorbers. (**E**) QFLS in the case of three bandgaps of perovskite with different HSLs. (**F**) Comparison of the V_{OC} of actual PSCs with the corresponding QFLS of representative layer stacks. (**G**) Comparison of the V_{OC} with different bandgaps of devices from the literature to our work. The line represents the V_{OC} extracted from calculated potential. The wide-bandgap region refers to 1.6 to 1.9 eV in the figure.

substrate (*29*, *30*), but few effective solutions to this problem have been explored. To estimate the thermal stability of the ITO/NiO_x/MeO-4PADBC substrate, we applied Kelvin probe force microscopy (KPFM) to record the surface potential evolution of the SAM under heat treatment by taking pristine ITO/MeO-4PADBC as a reference. Before thermal aging, the ITO/MeO-4PADBC and ITO/NiO_x/MeO-4PADBC substrates both exhibited relatively uniform surface potential with a narrow contact potential distribution (CPD) of ~40 mV, indicating that the SAM molecules were densely packed onto both of ITO and NiO_x surfaces (Fig. 3, A and B). After aging on a hotplate at 65°C for 1200 hours, the NiO_x/MeO-4PADBC substrate displayed negligible CPD changes, whereas the ITO/MeO-4PADBC CPD value increased to ~70 mV (Fig. 3, C and D). We propose that the fluctuations of CPD for the ITO/MeO-4PADBC substrate could be attributed to desorption, morphological changes under thermal stress, or both.

To further explore the binding ability of SAM to the substrate under heating, DFT simulations were conducted to investigate the binding energies between MeO-4PADBC and ITO or NiO_x substrate at 300 K (~27°C) and 340 K (~67°C) (Fig. 3, E and F). MeO-4PADBC had a higher binding energy with NiO_x (−22.4 eV) than with ITO (−16.7 eV) at 300 K, suggesting

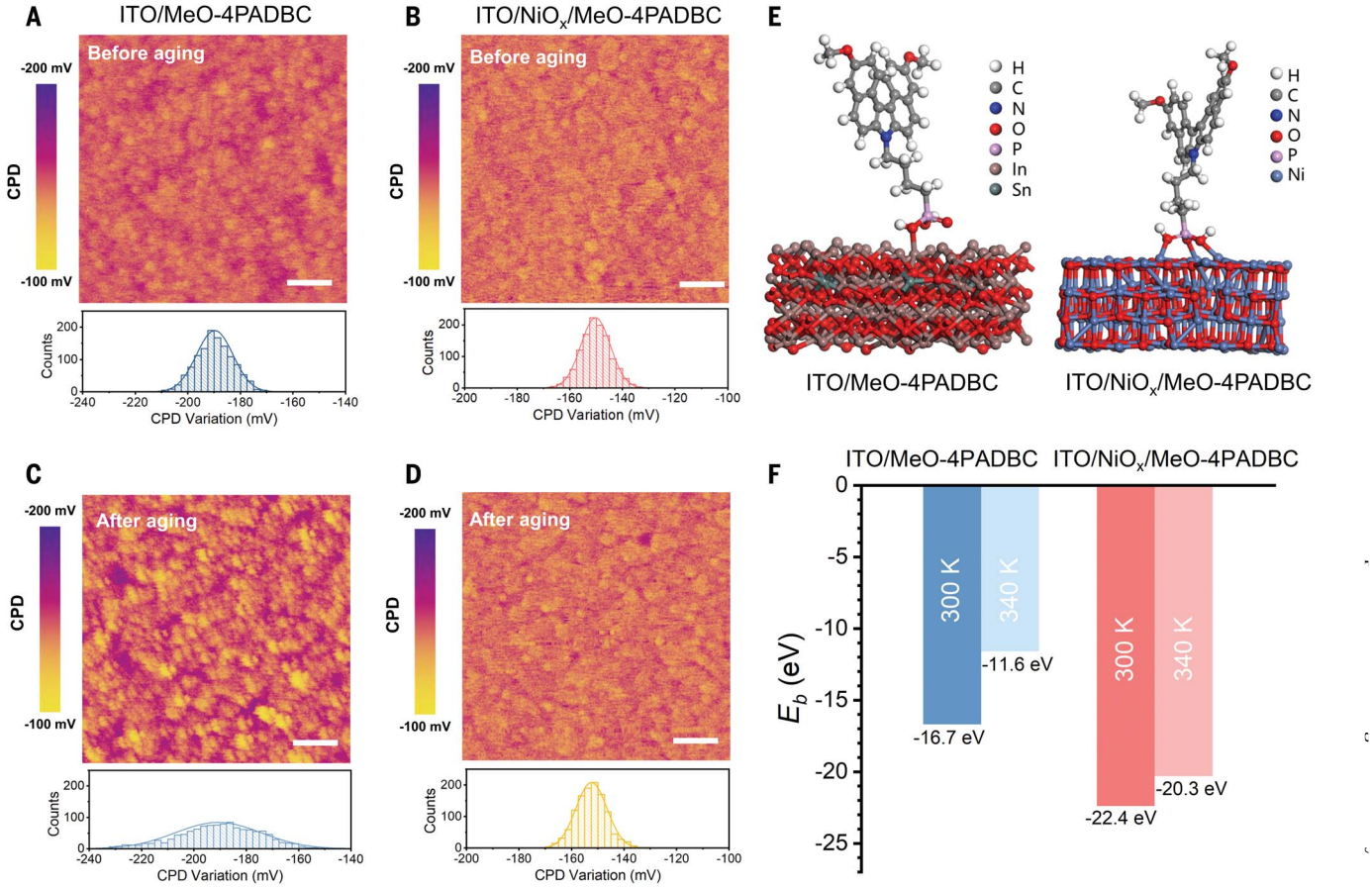


Fig. 3. Analysis of the degradation mechanism of PSCs. (**A** to **D**) Surface potential images obtained by scanning Kelvin probe microscopy (scale bar 500 nm) of the HSLs before and after aging at 65°C for 1200 hours. At the bottom of the figure are the statistical potential distributions of film surfaces. (**E**) DFT calculation of the binding energy of MeO-4PADBC with ITO and NiO_x. (**F**) Binding energies (E_b) of MeO-4PADBC with ITO and with NiO_x at the temperatures of 300 and 340 K.

a stronger bonding strength on the NiO_x film. This difference was attributed to the higher density of hydroxyl groups on the NiO_x surface than on the ITO, which is critical to the chemisorption of SAM on the metal oxide (*43*, *44*). The tridentate binding between SAM and NiO_x was stronger than the bidentate binding between SAM and ITO (*29*, *45*). At 340 K, the binding energy between SAM and ITO decreased from −16.7 to −11.6 eV, but that between MeO-4PADBC and NiO_x film showed minor changes (−20.3 eV). These results indicated that MeO-4PADBC on the NiO_x film is more robust against thermal stress compared with MeO-4PADBC on bare ITO.

Thermal accelerated aging measurements were conducted to evaluate the reliability of SAM-based PSCs. To avoid the influence of the 2D capping layer, the encapsulated devices without top passivation were prepared for stability testing. The initial *J-V* curves and photographs of the encapsulated devices are shown in fig. S40 and S41, respectively. The PSCs were operated under constant 1-sun illumination at fixed resistance loads near the MPP with the temperature ranging from 25° to 100°C, following the ISOS-L-2I procedure (*19*). The MeO-4PADBC-based devices degraded to 85 and 65% of starting PCEs after 1200 hours at 25°C and 65°C, respectively (Fig. 4A). When the aging temperature increased to 85°C, only 47% of its initial PCE was retained after 800 hours. The NiO_x devices retained 85 and 65% of initial PCEs at 65° and 85°C, respectively, after 1200 hours (Fig. 4B). However, the NiO_x/MeO-4PADBC-based devices retained 90 and 74% of initial PCEs after 1200 hours at 65° and 85°C, respectively (Fig. 4C), which was consistent with the degradation analysis in Fig. 3. Both devices based on NiO_x, MeO-4PADBC, or NiO_x/MeO-4PADBC as HSL showed a more obvious downward trend when the aging temperature was raised to 100°C. For a more comprehensive evaluation of the thermal stability of our NiO_x/MeO-4PADBC strategy, we further applied the PSCs with top passivation to conduct the operational stability at 65°C for 500 hours (fig. S42). The devices presented a slightly improved stability compared with the device without passivation at the same time stage, which maintained 96% of its initial PCE. This may be due to the higher hydrophobic surface brought by the CF_3-PEAI passivation (*46*).

We then determined the activation energy (E_a) of the temperature-dependent degradation of our PSCs with different HSLs according to a previously reported method (supplementary text) (*47*). The E_a value of NiO_x/MeO-4PADBC–based devices (0.389 ± 0.022 eV) was almost three times higher than that of MeO-4PADBC–based devices (0.150 ± 0.017 eV). The lifetime acceleration factor (AF) of each temperature can be obtained from E_a (Fig. 4D), from which it can be estimated that PSCs with NiO_x/MeO-4PADBC HSL could retained 80% of its initial PCE at room temperature after 7567 hours operation, without top passivation treatment.

Conclusions

We have demonstrated an efficient and stabilized HSL with greatly improved thermal stability for high efficiency SAM-containing inverted p-i-n PSCs. Rational molecular structure design of MeO-4PADBC and in-depth analysis revealed that optimal dipole moment and favorable contact with perovskite are the

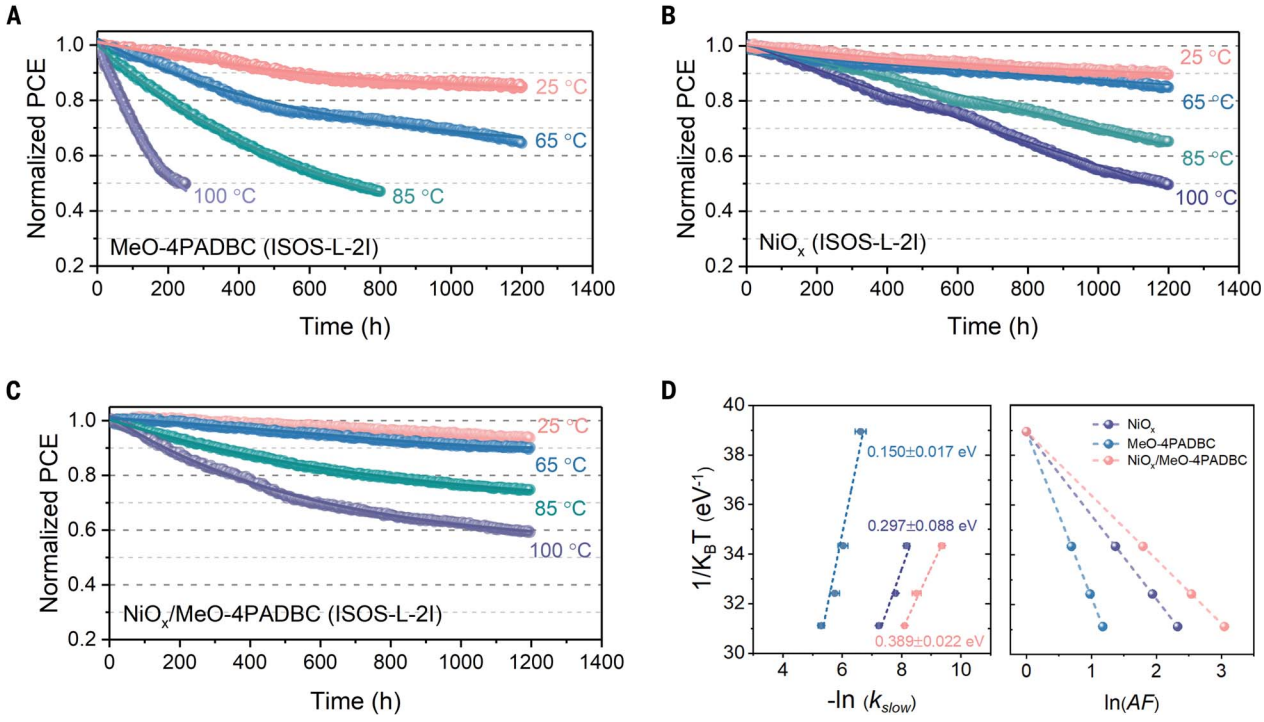


Fig. 4. Long-term stability assessment of PSCs under different temperatures. (**A** to **C**) Operational stability of PSCs with HSL of (A) MeO-4PADBC, (B) NiO_x, and (C) NiO_x/MeO-4PADBC at temperatures of 25°, 65°, 85°, and 100°C, respectively. (**D**) Natural logarithm of slow degradation rate (k_{slow}) versus $1/k_BT$ obtained from biexponential fits, where k_B is Boltzmann's constant and T is aging temperature. The dashed line comes from the linear fits to extract the E_a from each exponential (left); and natural logarithm of AF versus $1/k_BT$. Standard operating condition for AF value calculation refers to 1-sun illumination at 25°C (right).

keys to ideal energy alignment and fast hole extraction to improve the device efficiency and stability. Moreover, the anchoring of MeO-4PADBC SAM molecules on the NiO_x film can form a stronger tridentate bond with NiO_x, which effectively reduces the voltage loss and further maintains a strong fixation effect under thermal stress. Our study provides theoretical guidance for the design of efficient and stable HSL and paves the path for facile access to commercially available inverted p-i-n PSCs.

REFERENCES AND NOTES

1. Q. Jiang *et al.*, *Nature* **611**, 278–283 (2022).
2. Q. Tan *et al.*, *Nature* **620**, 545–551 (2023).
3. S. Zhang *et al.*, *Science* **380**, 404–409 (2023).
4. X. Zheng *et al.*, *Nat. Energy* **8**, 462–472 (2023).
5. H. Chen *et al.*, *Nature* **613**, 676–681 (2023).
6. X. Y. Chin *et al.*, *Science* **381**, 59–63 (2023).
7. S. Mariotti *et al.*, *Science* **381**, 63–69 (2023).
8. K. O. Brinkmann *et al.*, *Nature* **604**, 280–286 (2022).
9. R. He *et al.*, *Nature* **618**, 80–86 (2023).
10. A. Al-Ashouri *et al.*, *Energy Environ. Sci.* **12**, 3356–3369 (2019).
11. A. Al-Ashouri *et al.*, *Science* **370**, 1300–1309 (2020).
12. I. Levine *et al.*, *Joule* **5**, 2915–2933 (2021).
13. F. H. Isikgor *et al.*, *Nat. Rev. Mater.* **8**, 89–108 (2022).
14. C. M. Crudden *et al.*, *Nat. Chem.* **6**, 409–414 (2014).
15. L. Srisombat, A. C. Jamison, T. R. Lee, *Colloids Surf. A Physicochem. Eng. Asp.* **390**, 1–19 (2011).
16. H. Chen *et al.*, *Nat. Photonics* **16**, 352–358 (2022).
17. X. Li *et al.*, *Science* **375**, 434–437 (2022).
18. Z. Li *et al.*, *Science* **376**, 416–420 (2022).
19. M. V. Khenkin *et al.*, *Nat. Energy* **5**, 35–49 (2020).
20. S. M. Park *et al.*, *Science* **381**, 209–215 (2023).
21. G. Li *et al.*, *Science* **379**, 399–403 (2023).
22. G. Mani *et al.*, *Langmuir* **24**, 6774–6784 (2008).
23. P. Thissen, M. Valtiner, G. Grundmeier, *Langmuir* **26**, 156–164 (2010).
24. E. Delamarche, B. Michel, H. Kang, C. Gerber, *Langmuir* **10**, 4103–4108 (1994).
25. X. Wan *et al.*, *J. Phys. Chem. C* **124**, 2531–2542 (2020).
26. L. Li *et al.*, *Nat. Energy* **7**, 708–717 (2022).
27. W. Jiang *et al.*, *Angew. Chem. Int. Ed.* **61**, e202213560 (2022).
28. H. J. Yoon *et al.*, *Angew. Chem. Int. Ed.* **51**, 4658–4661 (2012).
29. J. Sun *et al.*, *Sol. RRL* **5**, 2100663 (2021).
30. E. Yalcin *et al.*, *Energy Environ. Sci.* **12**, 230–237 (2019).
31. R. Guo *et al.*, *Adv. Funct. Mater.* **33**, 2211955 (2023).
32. W. Wang *et al.*, *Adv. Energy Mater.* **13**, 2300694 (2023).
33. M. Li *et al.*, *Nat. Commun.* **14**, 573 (2023).
34. H. Zhang, C. Zhao, J. Yao, W. C. H. Choy, *Angew. Chem. Int. Ed.* **62**, e202219307 (2023).
35. S. Y. Kim, S. J. Cho, S. E. Byeon, X. He, H. J. Yoon, *Adv. Energy Mater.* **10**, 2002606 (2020).
36. I. Lange *et al.*, *Adv. Funct. Mater.* **24**, 7014–7024 (2014).
37. D. Di Girolamo *et al.*, *Chem. Sci.* **11**, 7746–7759 (2020).
38. F. Ma *et al.*, *J. Energy Chem.* **52**, 393–411 (2021).
39. Q. Lin, A. Armin, R. C. R. Nagiri, P. L. Burn, P. Meredith, *Nat. Photonics* **9**, 106–112 (2015).
40. J. Wang *et al.*, *Nat. Commun.* **11**, 177 (2020).
41. M. Stolterfoht *et al.*, *Adv. Mater.* **32**, e2000080 (2020).
42. M. Stolterfoht *et al.*, *Nat. Energy* **3**, 847–854 (2018).
43. N. Phung *et al.*, *ACS Appl. Mater. Interfaces* **14**, 2166–2176 (2022).
44. D. Koushik *et al.*, *J. Mater. Chem. C.* **7**, 12532–12543 (2019).
45. C. J. Flynn *et al.*, *ACS Appl. Mater. Interfaces* **8**, 4754–4761 (2016).
46. R. Chen *et al.*, *Nat. Energy* **8**, 839–849 (2023).
47. X. Zhao *et al.*, *Science* **377**, 307–310 (2022).

ACKNOWLEDGMENTS

Funding: This work was supported by the Innovation and Technology Fund (GHP/100/20SZ, GHP/102/20GD, MRP/040/21X), the ECS grant (21301319) and GRF grant (11306521) from the Research Grants Council of Hong Kong, Green Tech Fund (GTF202020164), the Science Technology and Innovation Committee of Shenzhen Municipality (SGDX20210823104002015, JCYJ20220818101018038). Zho.L. also acknowledges financial support from the National Natural Science Foundation of China (21975085) and Excellent Youth Foundation of Hubei Scientific Committee (2021CFA065). This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract DE-AC36-08GO28308. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. **Author contributions:** Zhe.L., X.S., X.Z., and B.L. contributed equally to this work. Z.Z. conceived the ideas and designed the project with Zho.L., J.M.L, and X.Z. Z.Z directed and supervised the research. Zhe.L. fabricated the devices, conducted the characterization, and analyzed the data. Zho.L. directed X.S. to design and synthesize the hole-selective molecule. B.L. and D.G. also contributed to the device fabrication, characterization, and data analyses. S.Z. conducted the DFT calculations. S.L. and J.G. also helped to fabricate the devices. Zhe.L., X.Z., B.L., Zho.L., J.M.L. and Z.Z. drafted, revised, and finalized the manuscript. All the authors revised the manuscript. **Competing interests:** Authors declare that they have no competing interests. The patent application been submitted. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.ade9637
Materials and Methods
Supplementary Text
Figs. S1 to S42
Tables S1 to S5
References (*48*–*67*)

Submitted 3 May 2023; accepted 6 September 2023
10.1126/science.ade9637

AVIAN ECOLOGY

Agriculture and hot temperatures interactively erode the nest success of habitat generalist birds across the United States

Katherine S. Lauck[1]*†, Alison Ke[1]†, Elissa M. Olimpi[1,2], Daniel Paredes[1,3], Kees Hood[1], Thomas Phillips[1], William R. L. Anderegg[4,5], Daniel S. Karp[1]

Habitat conversion and climate change are fundamental drivers of biodiversity loss worldwide but are often analyzed in isolation. We used a continental-scale, decades-long database of more than 150,000 bird nesting attempts to explore how extreme heat affects avian reproduction in forests, grasslands, and agricultural and developed areas across the US. We found that in forests, extreme heat increased nest success, but birds nesting in agricultural settings were much less likely to successfully fledge young when temperatures reached anomalously high levels. Species that build exposed cup nests and species of higher conservation concern were particularly vulnerable to maximum temperature anomalies in agricultural settings. Finally, future projections suggested that ongoing climate change may exacerbate the negative effects of habitat conversion on avian nesting success, thereby compromising conservation efforts in human-dominated landscapes.

Habitat conversion is the primary driver of terrestrial biodiversity loss, and climate change is projected to cause widespread extirpations (*1*, *2*). However, the effects of habitat conversion and climate change are often analyzed in isolation even though the fate of many species will ultimately depend on how they interact (*3*). For example, many forms of habitat conversion (e.g., agricultural or urban expansion) remove insulating tree canopies or other complex microhabitats, thereby exposing organisms to warmer maximum and/or cooler minimum temperatures [i.e., reducing thermal buffering (*4*)]. Indeed, temperatures in agricultural settings regularly attain levels >10°C higher than in nearby natural habitats (*5*). Other stressors related to human land use may increase the sensitivity of biodiversity to heat; for example, pesticide use and low vegetation complexity may reduce insect prey availability, limiting food and water available for thermoregulation (*6*). In addition, trees may protect understory species from heavy rains and retain moisture, buffering against drought (*4*). Thus, as temperatures warm and precipitation regimes shift, climate change may cause cities and farms to become even less hospitable, undermining efforts to safeguard biodiversity in human-dominated landscapes (*7*).

[1]Department of Wildlife, Fish, and Conservation Biology, University of California, Davis, Davis, CA, USA. [2]Department of Fish and Wildlife Conservation, Virginia Tech, Blacksburg, VA, USA. [3]Environmental Analysis Group, Department of Plant Biology, Ecology and Earth Science, University of Extremadura, Extremadura, Spain. [4]School of Biological Sciences, University of Utah, Salt Lake City, UT, USA. [5]Wilkes Center for Climate Science and Policy, University of Utah, Salt Lake City, UT, USA.
*Corresponding author. Email: kslauck@ucdavis.edu
†These authors contributed equally to this work.

Birds may be particularly sensitive to temperature extremes, because species with altricial young are ectothermic for the first few weeks of life, and extreme temperatures can divert energy from growth to thermoregulation (*8*). Although the effects of cold snaps on avian reproduction are well documented (*9*, *10*), recent work suggests that high temperatures can also reduce avian survival (*11*–*13*) and even cause community collapse (*13*). Temperature extremes may limit species persistence more than long-term increases in average temperatures (*14*), and variations in microclimate buffering can thus influence bird distributions (*15*).

Unfortunately, understanding the interactive effects of climate and land-use change on species persistence requires demographic data that are difficult to obtain over large spatiotemporal scales (*3*). Thus, our knowledge of how climate and land-use change interactively affect species is mostly restricted to changes in species distributions and/or abundances (*4*, *16*). Project NestWatch, a citizen-science nest-monitoring program organized by the Cornell Laboratory of Ornithology, offers a rare opportunity to explore how temperature extremes and land use interact to affect avian nesting success at a national scale. Using their data, we analyzed 152,863 nesting attempts by 58 bird species across 23 years (1998 to 2020) and 37,869 sites in four land-use types within the conterminous US: forests, open natural habitats, agricultural settings, and developed areas [table S1 (*17*)]. We used GridMET (*18*) to measure temperature extremes during each nesting attempt by calculating temperature anomalies, which we defined as the average maximum (or minimum) temperatures during the 45 days after each nesting attempt's date of first egg relative to conditions during the same date range over a historical reference period (1980 to 2000).

Our work was guided by four questi… First, how do the effects of temperature … nesting success vary across land-cover types? We hypothesized that maximum temperature anomalies would reduce success in open, human-dominated habitats with fewer thermally buffered areas. We also quantified precipitation anomalies during nesting, hypothesizing that extreme events would also reduce success in open, human-dominated habitats. Second, are some species more vulnerable to the interactive effects of habitat conversion and climate change than others? We predicted that species that build exposed nests (as opposed to cavity nests) and species of higher conservation concern would be the most sensitive to climate and land-use interactions. Third, are the effects of temperature and land cover consistent across species' ranges? We predicted that maximum temperature anomalies would disproportionately affect nesting success in agricultural areas within warmer regions. Finally, looking forward across the 21st century, how will nesting success likely change across space, time, and alternative climate change scenarios? We hypothesized that declines would be pronounced in agricultural settings, grassland, and developed areas, but not in forests.

The interactive effects of climate and land cover could arise from individuals of the same species varying in their temperature responses across land-cover types (e.g., if some provide more thermal buffering than others) or from shifts in species composition (e.g., if agriculture-associated species are more thermally sensitive than forest-associated species). To determine whether thermal buffering in some land-cover types increases avian resilience against heat waves, we focused on habitat generalist species that could conceivably exhibit different responses to heat waves in different land-cover types. Additionally, we modeled climate and land-use interactions using both generalized linear mixed models (GLMMs) that integrate within- and across-species effects and Bayesian models that only consider variation within species (*17*).

Climate and land-cover interactions

Interactions between maximum temperature anomalies and land-cover type

Across both modeling frameworks, we found that the effects of maximum temperature extremes on avian nesting success differed among land-use types ($P < 0.001$ for GLMM; Bayesian confidence interval: agricultural settings, −0.21 to −0.01; forests, 0.03 to 0.15; Fig. 1 and tables S2 to S5). In agricultural settings, the probability of successfully fledging at least one offspring declined by 46% (from 74 to 28%) between nests that experienced cooler versus warmer maximum temperature anomalies (i.e., 2 SDs lower versus 2 SDs higher than mean historical temperatures). Only considering

within-species effects reduced the decline from 46 to 30% over the same temperature range. In developed areas, the probability of success declined by only 5% from cooler to warmer maximum temperature anomalies (6% in the Bayesian model). One reason for the more muted decline might be that nests in "developed areas" were often in residential areas that can have high tree cover, not in city centers that may be subject to particularly extreme heat island effects. By contrast, only 1.5% of agricultural sites were in plantations with tree canopies (table S6). As a result, developed areas had higher canopy cover than both agricultural and natural open sites (mean cover: developed, 21%; agricultural, 11%; natural open, 17.6%). Future work could profitably focus on using temperature loggers to link nesting success to local microclimates rather than the coarser air temperature measurements analyzed here.

Although temperature increases in forests have been previously shown to reduce nest productivity through increased predation (*19*), we found that reproductive success increased in these areas by 14% across the same temperature range (21% in the Bayesian model). Why might this be? First, because tree canopies may keep nests cool, adults nesting in forests could be released from thermoregulatory care when temperatures increase, allowing them to spend more time foraging (*20*). Increased temperatures in cold microclimates might also reduce the reproductive cost of time spent off the nest (*21*). Maximum temperature anomalies could also drive phenological shifts that improve forest-nesting birds' access to food resources [e.g., by increasing insect abundance earlier in the season (*22*)]. Critically, the positive relationship between nesting success and maximum temperature anomalies in forests does not suggest that climate change is benefiting forest birds, because warming temperatures may be decreasing other demographic parameters (e.g., adult and/or juvenile survival).

Finally, in natural open habitats, nest success exhibited a nonlinear relationship with maximum temperature anomalies, peaking at an intermediate value (in the Bayesian model, no relationship with temperature was observed). One explanation is that grassland species could be adapted to temperature regimes in open-canopied environments, but too much deviation from historical norms (in either direction) may decrease success. Indeed, most nests in natural open habitats were in grasslands or prairies (~85% of attempts; table S6).

Our finding that the effects of temperature extremes on nesting success vary across land-use types persisted when additionally accounting for latitude, elevation, nest predation, spatial autocorrelation, extreme observations, the scale of landscape cover composition, and nests in more thermally buffered agricultural habitats [table S7 and fig. S1 (*17*)]. Birds were also not more sensitive to temperature anomalies in agricultural settings because of differences in air temperature among land-use types (fig. S2).

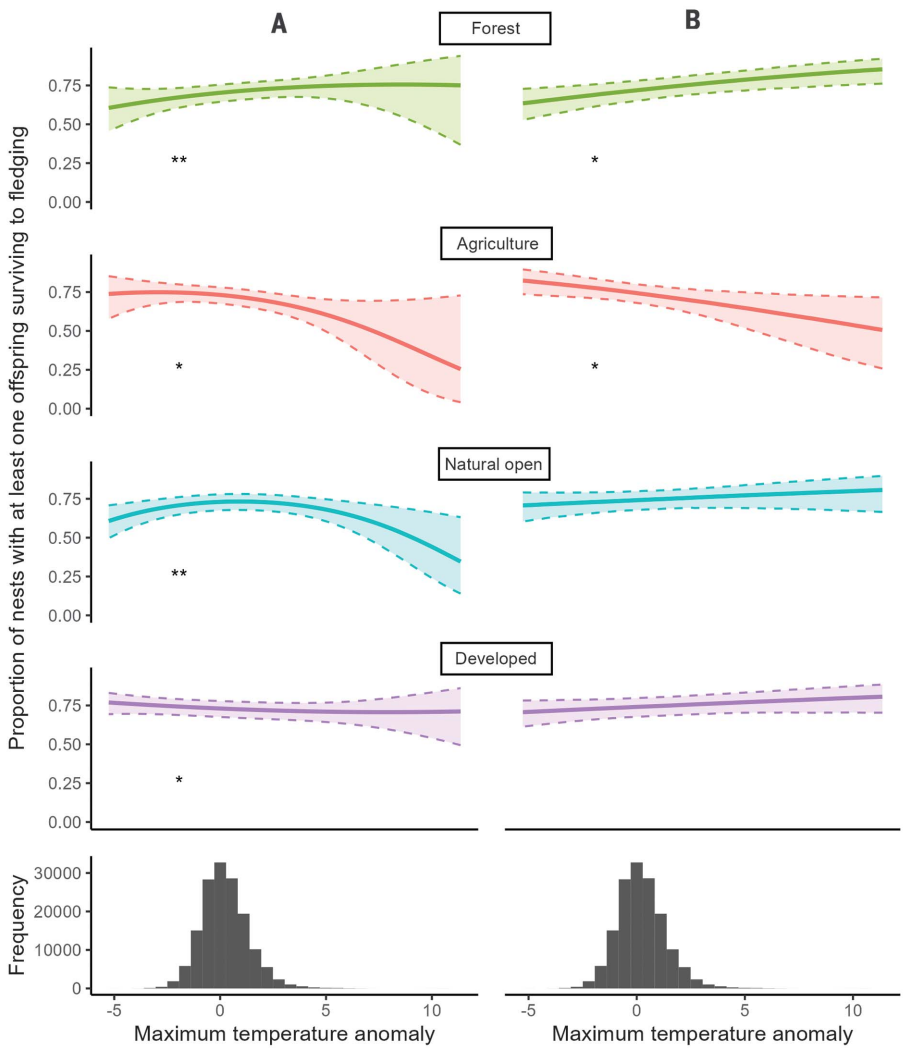


Fig. 1. Maximum temperature extremes reduce avian nesting success in agricultural settings but increase it in forests. For nests in each land-use type, panels present the relationship between maximum temperature anomalies 45 days after lay date as *z*-scores relative to historical temperatures (*17*) and the predicted proportion of nests with at least one offspring surviving to fledging. Solid lines indicate model predictions; shaded regions represent 95% confidence regions. Histograms in the bottom panels depict the distribution of maximum temperature anomalies across all nesting attempts. (**A**) Results from GLMMs combining within- and among-species variations in response to temperature. Asterisks indicate the level of significance in each land use. *$P < 0.05$; **$P < 0.01$. The quadratic effect of maximum temperature anomaly is only significant in natural open land cover. (**B**) Results from a Bayesian analysis that only assesses the effects of temperatures within species (i.e., factoring out changes in species composition among land-cover types). Asterisks indicate that the 95% Bayesian confidence interval did not overlap zero.

Interactions among minimum temperature anomaly, precipitation, and land cover

Unlike maximum temperatures, the effects of other climate variables exhibited much less variation among land-cover types (tables S2, S3, and S8). First, the linear effect of minimum temperature anomalies did not vary among land-cover types ($P = 0.28$). However, the quadratic effect did so marginally ($P = 0.05$), with nesting success exhibiting a more convex relationship with minimum temperature anomaly in forests and a more linear relationship in the other land-cover types. Second, precipitation over the prior year also exhibited no variation in linear effects on nesting success across land-cover types ($P = 0.19$). However, the quadratic effect again varied (table S3 and S8), exhibiting a more convex relationship in forests and a more linear effect in the other land-cover types. Finally, the linear effect of precipitation anomalies

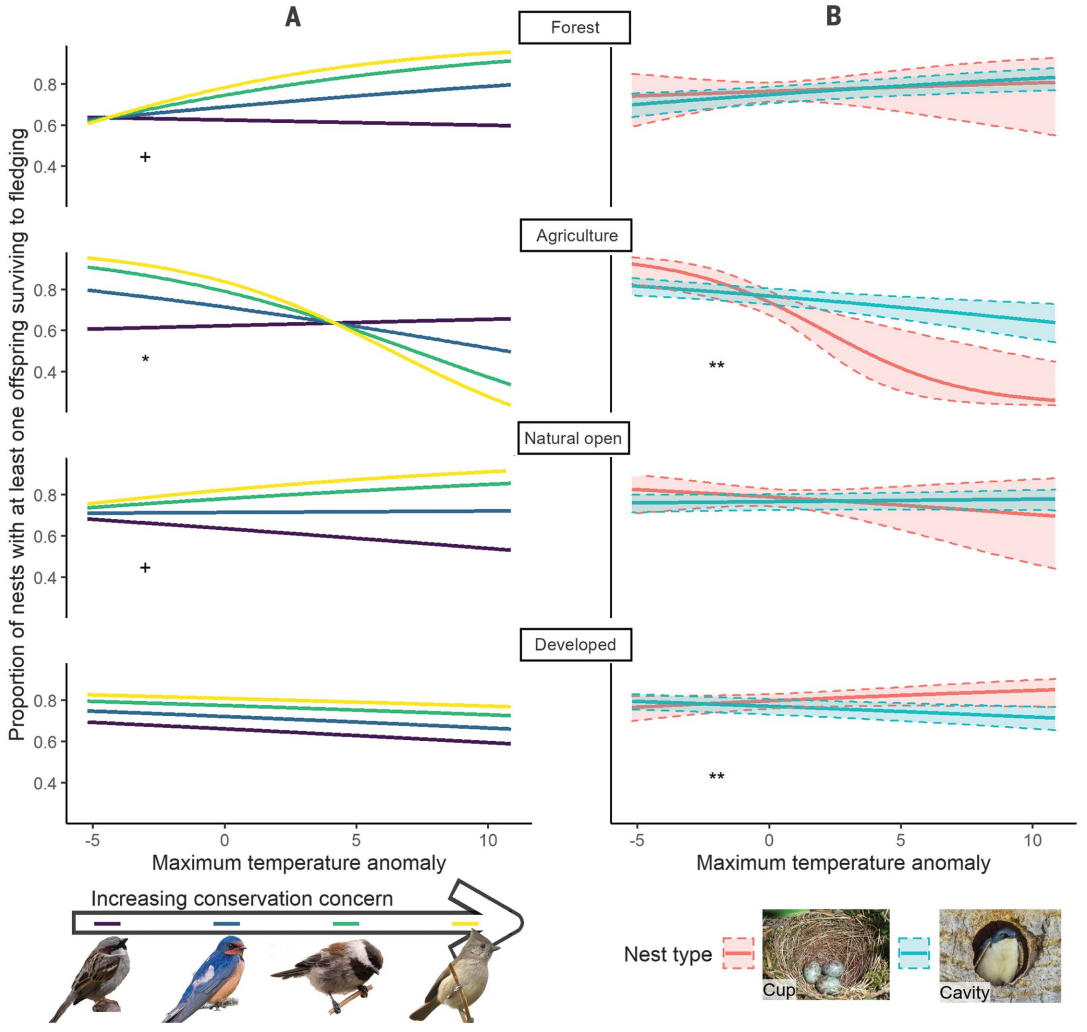


Fig. 2. Interactive effects of temperature extremes and land use are exacerbated for species of higher conservation concern and for cup-nesting species in agricultural settings. (**A**) Maximum temperature anomalies 45 days after lay date on the proportion of nests with at least one fledging (i.e., nest success) in each land-use type and for species of varying levels of conservation concern. Colored lines depict predictions for species of varying levels of conservation concern, represented by: house sparrow (*Passer domesticus*; score level 4, lowest concern), barn swallow (*Hirundo rustica*, level 8), chestnut-backed chickadee (*Poecile rufescens*, level 12), and oak titmouse (*Baeolophus inornatus*; score level 15, highest concern). (**B**) Maximum temperature anomaly effects on birds nesting in cavities (blue) versus cup nests (red). Asterisks indicate the level of significance of the interaction between maximum temperature anomalies and species' conservation scores or nest types in each land use. $+P < 0.1$, $*P < 0.05$, $**P < 0.01$.

during nesting varied by land-cover type, but with smaller effect sizes than maximum temperature anomalies ($P < 0.001$; fig. S3 and tables S3, S8, and S9). Specifically, higher precipitation during nesting was associated with higher success in agricultural settings, decreased success in forests, and small effects in other land-cover types. This suggests that birds in agricultural settings may suffer even more when precipitation is low and temperatures are high.

Identifying species sensitive to climate and land-cover interactions

The negative effects of maximum temperature anomalies on avian reproduction in agricultural settings and the positive effects in forests were generally consistent across species, but were not consistent across species in other land-cover types (fig. S4 and table S5). Multiple mechanisms may explain these trends. First, temperature extremes in agricultural settings could induce physiological stress and require nestlings to expend energy on thermoregulation rather than growth or to suffer water costs and dehydration (*8*). Second, maximum temperature anomalies may affect birds indirectly, for example, by reducing arthropod prey abundance or adult foraging efficiency in agricultural settings but not in natural habitats with microclimate refugia (*23*, *24*). Indeed, arthropods are particularly sensitive to climate warming in agricultural settings but are more buffered in natural habitats (*25*).

Nest type

Identifying traits that explain variations in species' responses to maximum temperature anomalies might provide insight into the underlying mechanisms. We found that cup nests in agricultural settings experienced particularly severe declines in nest success with higher maximum temperatures ($P = 0.005$; Fig. 2 and tables S10 to S12). Cup nests are less thermally buffered than cavity nests, suggesting that maximum temperature anomalies may indeed reduce nesting success directly [i.e., through avian physiology (*26*)]. We also found similar results when comparing nests in human-constructed nest boxes with other nests (fig. S5 and tables S10 to S12); however, because 99% of cavity nests were also in nest boxes, these two comparisons are functionally equivalent. This finding suggests that the effect sizes of the temperature and land-use interactions reported above may be conservative given the dominance of artificial nest box observations in the database (table S1).

Conservation concern

Species of higher conservation concern [as defined by (*27*)] were also more vulnerable to maximum temperature anomalies in agricultural settings and more successful in forests ($P = 0.027$; Fig. 2 and tables S10 to S12). This again suggests that the effect sizes reported above may be conservative for rarer, less frequently sampled species. For example, in agricultural settings, GLMMs predicted that hotter maximum temperature anomalies would decrease nest success by 69% (from 95 to 26%; maximum temperatures ± 2 SD) for species of highest concern, but would increase success by 5% for species of low concern. Because species of conservation concern are often already sensitive to anthropogenic land uses (*28*), they

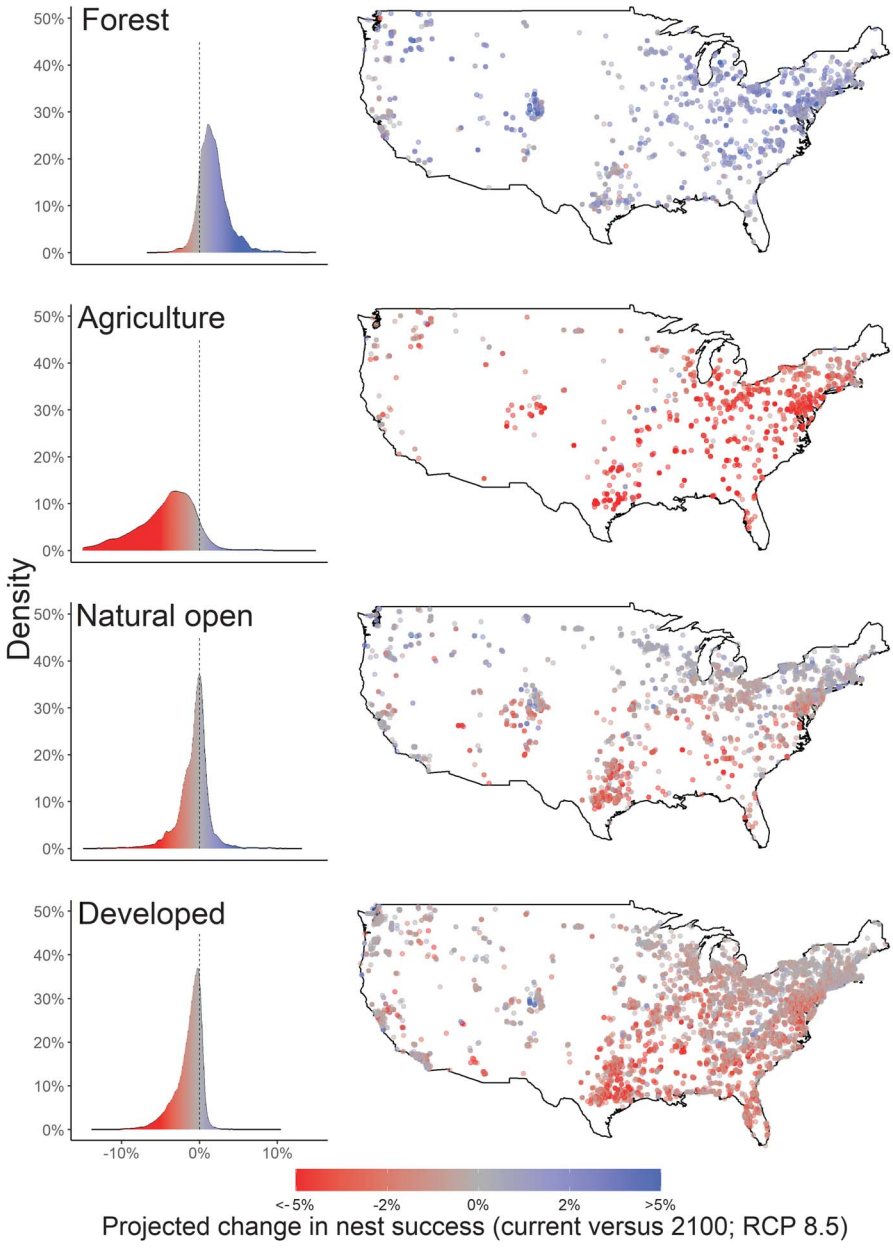


Fig. 3. Climate change is expected to decrease nesting success in agricultural settings but increase it in forests. Density diagrams (left) depict how the probability of successfully fledging at least one offspring is predicted to change by 2100 under a high-emissions scenario (RCP 8.5) across sites for each land-use type. Maps depict the spatial distribution of predicted changes; dark blue, red, and gray points indicate sites where nesting success is expected to increase by 5% or more, decrease by 5% or more, or remain largely unchanged, respectively. Points and histograms represent average predictions across five climate models (*17*).

may simply be unable to cope when temperatures increase in marginal environments.

Thermal limits

Avian reproduction is expected to be most sensitive to temperature extremes in regions that experience temperatures near species' thermal limits (*29*). To examine spatial variation in species' responses to maximum temperature anomalies and land use, we calculated site temperature baselines in four ways, including approaches that quantify temperatures in absolute terms and relative to temperature variability across each species' geographic range (*17*). Although nest success was more sensitive to maximum temperature anomalies in warmer regions ($P < 0.001$ for three of four measures; fig. S6 and tables S13 and S14), the interactive effect of maximum temperature anomalies and land-use types was consistent across regions: In both hot and cold regions, maximum temperature anomalies were the most harmful in agricultural settings ($P \geq 0.6$; tables S13 and S15).

Future climate change scenarios

Our results suggest that bird species nesting in agricultural settings may be more vulnerable to climate change than those in forests. Leveraging five global climate models and multiple climate change scenarios [representative concentration pathways (RCPs)], we used our models to explore how nest success would change if each nesting attempt in our dataset instead occurred with the climate conditions of 2040 to 2059 or 2080 to 2099 (Fig. 3, fig. S7A, and table S16). These analyses implicitly assume that nesting phenology is fixed in time, even though some birds can shift to earlier breeding times to track climate niches (*29*). Therefore, these results should be considered a sensitivity analysis rather than explicit predictions. Climatic uncertainty for agricultural settings was slightly higher than for the other land-cover types, especially in the southeastern US (figs. S7B and S8), but we suspect that climatic uncertainty may have constrained the projected effect sizes. Statistical uncertainty exceeded climatic uncertainty (figs. S7C and S8). Our models suggest that nesting success in agricultural settings would decline by an additional 4.41% by 2100 if greenhouse gas emissions maintain their current rate of increase (RCP 8.5), whereas success in forests would increase by 1.78%. By contrast, if emissions were reduced (RCP 4.5), then nesting success in agricultural settings would decline by only 1.14% and success in forests would increase by 1.07% (fig. S7A and table S16). Thus, if emissions are curtailed, then birds nesting in human-dominated areas would likely be more successful over the long term.

Conclusions

Our results highlight the vulnerability of birds nesting in agricultural settings to temperature extremes and may offer insight into mechanisms underlying North American bird declines (*30*). They also align with recent findings from Europe suggesting that climate change may be causing larger population declines in generalist, farmland-associated birds compared with specialist, woodland-associated species (*31*). Maintaining forest patches in anthropogenic landscapes may thus increase avian resilience to extreme climatic events (*32*). An important caveat is that the species studied here are habitat generalists. Enhancing forest cover in naturally open land covers could harm grassland-obligate birds, which might be able to leverage habitat heterogeneity in natural grasslands to find thermally buffered areas (*33*). Nonetheless, for other species, erecting

sun-shielded or insulated nest boxes in shaded locations (*34*), conserving forest patches, and/or planting scattered trees in human-dominated landscapes may help species cope with climate change–driven temperature extremes by providing thermal buffering (*35*), especially for the species of conservation concern studied here.

REFERENCES AND NOTES

1. R. Dirzo, P. H. Raven, *Annu. Rev. Environ. Resour.* **28**, 137–167 (2003).
2. F. E. B. Spooner, R. G. Pearson, R. Freeman, *Glob. Chang. Biol.* **24**, 4521–4531 (2018).
3. T. H. Oliver, M. D. Morecroft, *Wiley Interdiscip. Rev. Clim. Change* **5**, 317–335 (2014).
4. J. J. Williams, T. Newbold, *Divers. Distrib.* **26**, 76–92 (2020).
5. R. A. Senior, J. K. Hill, P. González Del Pliego, L. K. Goode, D. P. Edwards, *Ecol. Evol.* **7**, 7897–7908 (2017).
6. D. R. Garrett, F. Pelletier, D. Garant, M. Bélisle, *Ecol. Monogr.* **92**, e1518 (2022).
7. C. Kremen, A. M. Merenlender, *Science* **362**, eaau6020 (2018).
8. J. C. Wingfield *et al*., *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, 20160140 (2017).
9. D. W. Winkler, M. K. Luo, E. Rakhimberdiev, *Oecologia* **173**, 129–138 (2013).
10. J. R. Shipley *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **117**, 25590–25594 (2020).
11. B. Zuckerberg, C. A. Ribic, L. A. McCauley, *Conserv. Biol.* **32**, 872–882 (2018).
12. A. R. Bourne, S. J. Cunningham, C. N. Spottiswoode, A. R. Ridley, *Proc. Biol. Sci.* **287**, 20201140 (2020).
13. E. A. Riddell, K. J. Iknayan, B. O. Wolf, B. Sinervo, S. R. Beissinger, *Proc. Natl. Acad. Sci. U.S.A.* **116**, 21609–21615 (2019).
14. N. E. Zimmermann *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19723–19728 (2009).
15. S. J. K. Frey, A. S. Hadley, M. G. Betts, *Divers. Distrib.* **22**, 944–959 (2016).
16. J. M. Northrup, J. W. Rivers, Z. Yang, M. G. Betts, *Glob. Chang. Biol.* **25**, 1561–1575 (2019).
17. Materials and methods are available as supplementary materials.
18. J. T. Abatzoglou, *Int. J. Climatol.* **33**, 121–131 (2013).
19. W. A. Cox, F. R. Thompson 3rd, J. L. Reidy, J. Faaborg, *Glob. Chang. Biol.* **19**, 1064–1074 (2013).
20. V. G. Rohwer, J. R. Purcell, *PLOS ONE* **14**, e0219907 (2019).
21. B. H. Coe, M. L. Beck, S. Y. Chin, C. M. B. Jachowski, W. A. Hopkins, *J. Avian Biol.* **46**, 385–394 (2015).
22. E. K. Bowers *et al*., *Ecology* **97**, 2880–2891 (2016).
23. A. J. Suggitt *et al*., *Oikos* **120**, 1–8 (2011).
24. C. Funghi, L. S. C. McCowan, W. Schuett, S. C. Griffith, *Anim. Behav.* **149**, 33–43 (2019).
25. C. L. Outhwaite, P. McCann, T. Newbold, *Nature* **605**, 97–102 (2022).
26. M. C. Mainwaring, I. R. Hartley, M. M. Lambrechts, D. C. Deeming, *Ecol. Evol.* **4**, 3909–3928 (2014).
27. North American Bird Conservation Initiative, "The state of North America's birds 2016," in *Environment and Climate Change* (Cornell Univ., 2016); www.stateofthebirds.org.
28. L. Sykes, L. Santini, A. Etard, T. Newbold, *Conserv. Biol.* **34**, 688–696 (2020).
29. J. B. Socolar, P. N. Epanchin, S. R. Beissinger, M. W. Tingley, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 12976–12981 (2017).
30. K. V. Rosenberg *et al*., *Science* **366**, 120–124 (2019).
31. S. Rigal *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2216573120 (2023).
32. D. G. Nimmo, A. Haslem, J. Q. Radford, M. Hall, A. F. Bennett, *J. Appl. Ecol.* **53**, 449–458 (2016).
33. J. S. Bernath-Plaisted, C. A. Ribic, W. B. Hills, P. A. Townsend, B. Zuckerberg, *Environ. Res. Lett.* **18**, 064023 (2023).
34. E. R. Larson, J. R. Eastwood, K. L. Buchanan, A. T. D. Bennett, M. L. Berg, *Ecol. Manage. Restor.* **19**, 39–48 (2018).
35. T. Vanwalleghem, R. K. Meentemeyer, *Ecosystems* **12**, 1158–1172 (2009).
36. Data for: K. S. Lauck *et al*., Agriculture and hot temperatures interactively erode the nest success of habitat generalist birds across the United States, Dryad (2023); https://doi.org/10.25338/B8ZD1P.
37. Code for: K. S. Lauck *et al*., Agriculture and hot temperatures interactively erode the nest success of habitat generalist birds across the United States, Zenodo (2023); https://doi.org/10.5281/zenodo.8226606.

ACKNOWLEDGMENTS

We thank the volunteers who contributed data to the Cornell Laboratory of Ornithology's Project NestWatch program, as well as program staff; R. Bailey for help in preparing and sharing the database; W. Brooks of the UC Davis DataLab for invaluable advice regarding statistical approaches for spatial autocorrelation; and our anonymous reviewers for their rigorous and constructive feedback. **Funding:** This work was supported by a UC Davis Graduate Group in Ecology Fellowship (K.S.L.); a National Science Foundation Graduate Research Fellowship (A.K.); and an Achievement Rewards for College Scientists Fellowship (A.K.).
Author contributions: Conceptualization: K.S.L., A.K., E.M.O., D.P., K.H., T.P., W.R.L.A., D.S.K.; Data curation: K.S.L., A.K., E.M.O., D.P., W.R.L.A., D.S.K.; Formal analysis: K.S.L., A.K., E.M.O., D.P.; Funding acquisition: K.S.L., A.K., D.S.K.; Investigation: K.S.L., A.K., E.M.O., D.P., K.H., T.P., W.R.L.A., D.S.K.; Methodology: K.S.L., A.K., E.M.O., D.P., K.H., T.P., W.R.L.A., D.S.K.; Supervision: D.S.K.; Visualization: K.S.L., A.K., E.M.O., D.P.; Writing – original draft: K.S.L., A.K., D.S.K.; Writing – review and editing: K.S.L., A.K., E.M.O., D.P., K.H., T.P., W.R.L.A., D.S.K. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript, the supplementary materials, or Dryad (*36*). Code are available on Zenodo (*37*). Data and codebase are available under the GNU General Purpose License 3.0. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.add2915
Materials and Methods
Figs. S1 to S8
Tables S1 to S16
References (*38*–*49*)

Submitted 1 June 2022; accepted 11 August 2023
10.1126/science.add2915

RADIO ASTRONOMY

A luminous fast radio burst that probes the Universe at redshift 1

S. D. Ryder[1,2], K. W. Bannister[3], S. Bhandari[4,5], A. T. Deller[6], R. D. Ekers[3,7], M. Glowacki[7], A. C. Gordon[8], K. Gourdji[6], C. W. James[7], C. D. Kilpatrick[8,9], W. Lu[10,11], L. Marnoch[1,2,3,12], V. A. Moss[3], J. X. Prochaska[13,14], H. Qiu[15], E. M. Sadler[3,16], S. Simha[13], M. W. Sammons[7], D. R. Scott[7], N. Tejos[17], R. M. Shannon[6]*

Fast radio bursts (FRBs) are millisecond-duration pulses of radio emission originating from extragalactic distances. Radio dispersion is imparted on each burst by intervening plasma, mostly located in the intergalactic medium. In this work, we observe the burst FRB 20220610A and localize it to a morphologically complex host galaxy system at redshift 1.016 ± 0.002. The burst redshift and dispersion measure are consistent with passage through a substantial column of plasma in the intergalactic medium and extend the relationship between those quantities measured at lower redshift. The burst shows evidence for passage through additional turbulent magnetized plasma, potentially associated with the host galaxy. We use the burst energy of 2×10^{42} erg to revise the empirical maximum energy of an FRB.

Fast radio bursts (FRBs) (*1*, *2*) are transient radio sources that last a few milliseconds, emitted by extragalactic sources. Free electrons along the path between the FRB source and Earth impart a frequency-dependent time delay (dispersion) on the radio signal. This dispersion can be used to measure the column density of free electrons [quantified by the dispersion measure (DM)] between the FRB source and the observer. FRBs localized to host galaxies at different redshifts exhibit a positive correlation between the extragalactic DM and host redshift, known as the Macquart relation (*3*). This relation has been used to measure the cosmic baryon fraction and the expansion rate of the Universe (*4*). This relation has been measured using identified FRB host galaxies at nearby redshifts of $z \lesssim 0.5$. Some unlocalized FRBs (with unknown host galaxies) have DMs consistent with $z > 1$ (*5*); however, an FRB associated with a galaxy at $z = 0.241$ had a high DM that would have implied $z > 1$ if its host had been unknown (*6*). This demonstrates that estimates of redshift from DM alone can be misleading if they are affected by plasma within the host galaxy—which also imparts a contribution to the DM—in addition to that of the intergalactic medium.

[1]School of Mathematical and Physical Sciences, Macquarie University, Sydney, NSW 2109, Australia. [2]Astrophysics and Space Technologies Research Centre, Macquarie University, Sydney, NSW 2109, Australia. [3]Australia Telescope National Facility, Commonwealth Science and Industrial Research Organisation, Space and Astronomy, Epping, NSW 1710, Australia. [4]Netherlands Institute for Radio Astronomy (ASTRON), 7991 PD Dwingeloo, Netherlands. [5]Joint institute for Very Long Baseline Interferometry in Europe, 7991 PD Dwingeloo, Netherlands. [6]Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Hawthorn, VIC 3122, Australia. [7]International Centre for Radio Astronomy Research, Curtin Institute of Radio Astronomy, Curtin University, Perth, WA 6102, Australia. [8]Center for Interdisciplinary Exploration and Research in Astrophysics, Northwestern University, Evanston, IL 60208, USA. [9]Department of Physics and Astronomy, Northwestern University, Evanston, IL 60208, USA. [10]Department of Astronomy University of California, Berkeley, CA 94720, USA. [11]Theoretical Astrophysics Center, University of California, Berkeley, CA 94720, USA. [12]Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), Macquarie University, Sydney, NSW 2109, Australia. [13]Department of Astronomy and Astrophysics, University of California, Santa Cruz, CA 95064, USA. [14]Kavli Institute for the Physics and Mathematics of the Universe, Kashiwa, 277-8583, Japan. [15]Square Kilometre Array Observatory, Jodrell Bank, Lower Withington, Macclesfield SK11 9FT, UK. [16]Sydney Institute for Astronomy, School of Physics, University of Sydney, Sydney, NSW 2006, Australia. [17]Instituto de Física, Pontificia Universidad Católica de Valparaíso, Casilla 4059, Valparaíso, Chile.
*Corresponding author. Email: rshannon@swin.edu.au

Observations of FRB 20220610A

FRBs have been searched for (*7*) and localized (*8*) using the Australian Square Kilometre Array Pathfinder (ASKAP) (*9*), a radio interferometer in Western Australia comprising 36 antennas, each 12 m in diameter. Each antenna is equipped with a phased array receiving system, which provides 36 beams across the focal plane, covering ~30 square degrees. FRBs are searched for in real time using the incoherent sum of the intensities of each antenna in each beam. When an FRB is detected, temporary voltage buffers are downloaded, correlated, calibrated, and imaged; this enables the positions of detected bursts to be measured with an absolute precision of typically a few tenths of an arc second (*10*, *11*).

We detected FRB 20220610A in observations taken around the position of the previously known burst FRB 20220501C, but the two bursts are not related (*11*). The observations were centered at a frequency of 1271.5 MHz and had a time resolution of 1.18 ms. The dispersion of FRB 20220610A indicated a DM of $1458.15^{+0.25}_{-0.55}$ pc cm^{-3} (where pc is parsecs). This is higher than all but one of the 55 FRBs that were previously observed using ASKAP (*4*). A dedispersed dynamic spectrum of the burst is shown in Fig. 1, and observed properties of the burst are listed in Table 1. The burst does not show the 10- to 100-MHz modulation in

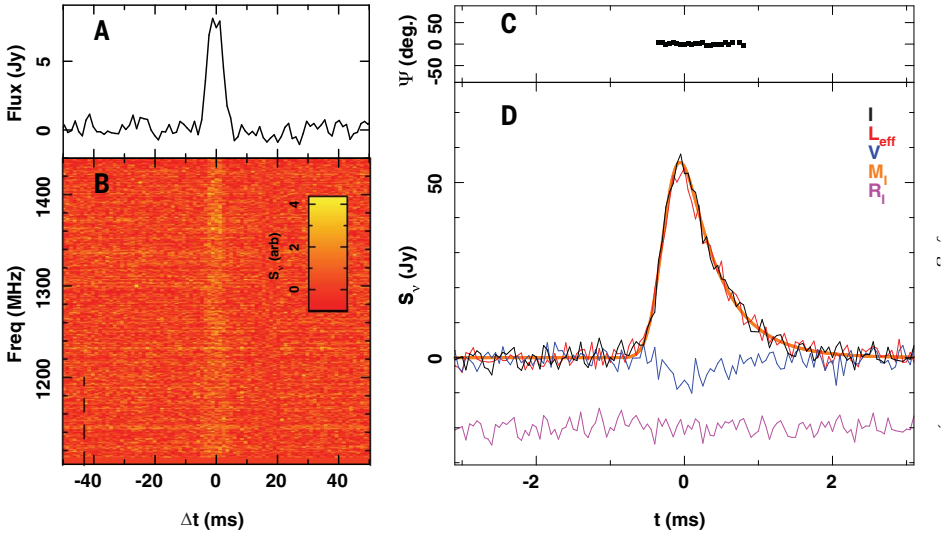


Fig. 1. Radio observations of FRB 20220610A. (**A** and **B**) The burst as observed in the incoherently summed data stream used by the real-time detection system. The pulse width is dominated by intrachannel dispersion smearing. (A) The integrated flux density (S_ν) of the pulse profile as a function of time (t). (B) The dedispersed burst dynamic spectrum as a function of frequency (Freq). The vertical dashed line in (B) indicates the frequency range (<1200 MHz) of high–time resolution data that were saved by the pipeline; data at higher frequencies were lost as a result of latency issues (see the text for details). (**C** and **D**) High–time resolution data (frequencies <1200 MHz) produced from the raw electromagnetic field samples saved from each telescope after the real-time trigger (*11*). (C) Linear polarization position angle Ψ during the pulse. (D) The integrated pulse profile in total intensity I (black), linear polarization $L_{\rm eff}$ (red), Stokes V (blue), and a maximum likelihood model of Stokes I, denoted M_I (orange). The residuals R_I (pink) between I and M_I are shown offset from zero by −20 Jy for clarity. The polarization position angle Ψ is defined such that linear polarization is in one Stokes parameter and hence $\Psi \approx 0$.

the spectrum characteristic of many lower-DM, high–Galactic latitude ASKAP-detected FRBs (*7*).

The 2-s dispersive sweep of the burst across the instrument bandwidth and the 2.4-s latency in the detection system resulted in only the lowest 88 MHz of the burst being captured in the 3.1-s-duration voltage buffer. This was sufficient to localize the burst to a precision of 0.5 arc sec. We used the voltage data to reconstruct the high time resolution and polarimetric properties of the burst (*11*). After correcting for dispersive smearing, the burst shows an exponentially decreasing tail (Fig. 1D), which is consistent with scatter broadening as a result of turbulence in intervening plasma (*12*). We measure the pulse broadening time to be 0.511 ± 0.012 ms at a reference frequency of 1147.5 MHz, assuming a ν^{-4} frequency dependence (where ν is the frequency).

Ordered magnetic fields in astrophysical plasmas add additional, polarization-dependent dispersion. This manifests as wavelength-dependent variation in the linear polarization position angle, referred to as Faraday rotation (*13*). FRB 20220610A exhibits Faraday rotation, with a rotation measure (RM) of 215 ± 2 rad m^{-2}. After correcting for this Faraday rotation, we calculate that the burst had an intrinsic linear polarization fraction of 96 ± 1%. The high fractional linear polarization allows us to place a 67% upper limit on the Faraday dispersion (σ_{RM} < 0.6 rad m^{-2}) induced by fluctuations in the RM in intervening turbulent plasma. Higher levels of Faraday dispersion have been detected for other FRBs (*14*). The burst also shows lower fractional circular polarization of 10 ± 1%. Although instrumental artifacts can induce spurious circular polarization, we do not see any correlation between the Stokes polarization parameters *U* and *V* in the spectrum, which would be expected for an instrumental effect (*11*). The FRB was located ~4 arc min from the beam center, which makes off-axis leakage effects unlikely (*15*). Circular polarization has been observed in some FRBs and could either be intrinsic to the burst (*16*) or result from propagation through relativistic plasma in the immediate source environment (*17*).

Table 1. Properties of FRB 20220610A and its host galaxy. The fluence was derived from filter bank data that include the full ASKAP bandwidth. UTC, coordinated universal time; S/N, signal-to-noise ratio; FWHM, full width at half maximum. The host photometry is expressed in magnitudes (mag) on the AB system. The star formation rate (SFR) is in solar masses per year ($M_\odot$ year^{-1}). 100 Myr SFR is the integrated star formation in the past 100 million years (Myr). *Z*, metallicity of the host galaxy; $Z_\odot$, metallicity of the Sun.

Measured burst properties	
DM	$1458.15^{+0.25}_{-0.55}$ pc cm^{-3}
Topocentric arrival time at 1104 MHz (UTC)	2022-06-10 22:26:44.313
Fluence	45 ± 5 Jy ms
Right ascension (J2000 equinox)	23^{h}24^{m}17^s.569 ± 0^s.040
Declination (J2000 equinox)	−33°30′49″.37 ± 0″.50
Galactic longitude	8.83954°
Galactic latitude	−70.18569°
Incoherent detection S/N (1104–1440 MHz)	29.8
Image S/N (1104–1152 MHz)	81
RM	215 ± 2 rad m^{-2}
Inferred burst properties	
Intrinsic width (FWHM)	0.41 ± 0.01 ms
Implied FRB isotropic energy density	6 × 10^{32} erg Hz^{-1}
Milky Way disk DM contribution	31 pc cm^{-3}
Measured host galaxy properties	
Redshift	1.016 ± 0.002
Photometry:	
Band (central wavelength, Å)	Magnitude (mag)
g (4700)	24.15 ± 0.07
V (5510)	23.89 ± 0.13
R (6580)	23.78 ± 0.06
I (8060)	22.17 ± 0.07
z (9620)	21.95 ± 0.12
J (12,200)	21.97 ± 0.07
K_s (21,460)	22.08 ± 0.12
Inferred host galaxy properties	
Mass-weighted age	$1.02^{+1.64}_{-0.62}$ Gyr
log(stellar mass)	$9.98^{+0.14}_{-0.07}$ $M_\odot$
log(total mass)	$10.11^{+0.18}_{-0.07}$ $M_\odot$
100 Myr SFR	$0.42^{+0.83}_{-0.37}$ $M_\odot$ year^{-1}
log($Z/Z_\odot$)	$-0.11^{+0.17}_{-1.68}$

Host galaxy properties

We performed follow-up ground-based optical and infrared observations with the Very Large Telescope (VLT) and the W. M. Keck Observatory to identify and characterize the host galaxy of FRB 20220610A (*11*). The images (Fig. 2, A to C) show an object at the FRB location that has an extended, multicomponent morphology. We label the optical source that overlaps the radio position of the FRB as component (a) and two adjacent sources as components (b) and (c) (Fig. 2A). We use a Bayesian method to assess the chance of coincidence between a radio transient and optical galaxies (*18*), finding >99.99% confidence that the FRB is associated with component (a).

We performed broad-band optical and infrared spectroscopy of components (a), (b), and (c) (Fig. 2, D and E) (*11*). We identify two emission lines in the spectra as the [O II] 3726 and 3729 Å doublet, which is most prominent in the spectrum of component (b) (fig. S2). We measure the redshift of each component from this doublet, finding that they are all consistent with z = 1.016 ± 0.002.

We use the photometry and a stellar population model to estimate the total mass of the three components combined as 10^{10} solar masses, with a star formation rate of ~0.42 solar masses per year (*11*). These values, in addition to the estimated host metallicity and star formation history, are consistent with those of nearby FRB hosts (*19*, *20*), but the source morphology is markedly different. Observed and estimated properties of the host galaxy are also listed in Table 1.

The presence of two bright components (a) and (c) separated by 2.0 arc sec (which corresponds to a distance of 16 kpc at that redshift), and the diffuse feature (b) between them, is consistent with two galaxies interacting or merging or with a compact galaxy group. It is also possible that the morphology is due to

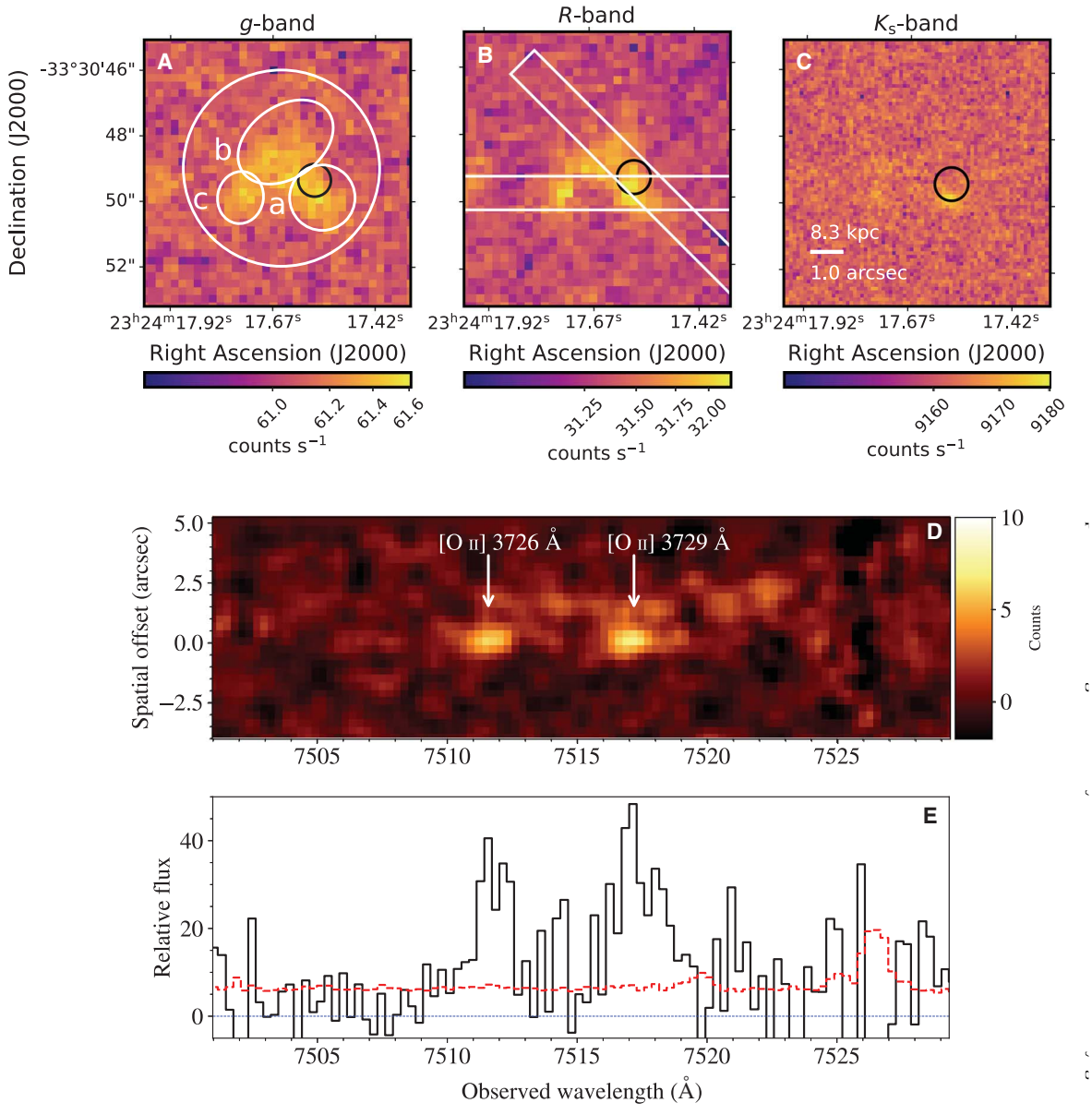


Fig. 2. Optical and infrared observations of the host galaxy of FRB 20220610A. In (A) to (C), the FRB localization and its uncertainty (68% confidence) are indicated by the black ellipse. (**A**) Optical *g*-band VLT image. White ellipses outline the apertures used for the photometry of each component (table S2). The larger unlabeled white circle is the aperture used for the entire combined system. (**B**) Optical *R*-band image. White rectangles outline the slit locations used for VLT spectroscopy. (**C**) Infrared K_S-band image. The scale bar applies to (A) to (C); it shows the angular scale and corresponding projected physical scale at the measured host redshift. The color bar below each panel indicates relative count rates in each processed image. (**D**) Two-dimensional (2D) VLT spectroscopy at a slit position angle on the sky of 45° covering components (a) and (b), as marked in (B). White arrows indicate two lines due to the [O II] doublet, with rest wavelengths of 3726 and 3729 Å. (**E**) 1D spectrum (black solid line) and its uncertainty (red dashed line) extracted from the 2D spectrum in (D), centered on the peak of the [O II] lines using an aperture width of 1.8″. The blue line indicates zero relative flux in arbitrary units.

internal structure within a single galaxy; at these redshifts, about half of all galaxies have clumpy morphologies (*21*). We regard the latter possibility as unlikely because of the large spatial separation between the components. Only component (a) is detected in the near-infrared (K_s-band) image (Fig. 2C), which indicates that it hosts an older stellar population compared with the other components. Component (a) is also displaced from the centroid of the total optical light in the *g*- and *R*-bands, contrary to what would be expected if it was the nuclear bulge of a single galaxy.

Extending the Macquart relation

We used the measured properties of FRB 20220610A to investigate the Macquart relation out to $z \sim 1$, by comparing the observed DM with predictions based on previous measurements of the relation at $z \leq 0.522$. Figure 3 shows the relationship between DM and redshift for the FRBs detected by ASKAP (*4*). We restrict our analysis to the ASKAP sample to reduce the effect of selection effects, which depend on the observing system. We do not refit the Macquart relation because doing so in an unbiased way would require a reanalysis of the entire updated FRB sample from ASKAP.

After subtracting a model Milky Way foreground contribution to the observed DM (*11*), we estimate the non-Galactic DM of FRB 20220610A to be ≈1376 pc cm^{-3}, indicating a high column density of ionized gas between the FRB and Earth. This is higher than the DM expected from the Macquart relation, by ~650 pc cm^{-3}, which is a 2.4σ excess (*11*). If the excess is real and originates from the host galaxy, the implied electron column density is 1300^{+170}_{-320} pc cm^{-3} in the host rest frame (*11*), with the uncertainty reflecting the intrinsic variation in the contribution from the intergalactic medium.

Interpretation of the DM

We use the scatter broadening and Faraday rotation of FRB 20220610A to investigate the properties of plasma at $z \approx 1$. A dispersion excess could potentially arise from any combination of gas in the immediate vicinity of the source, the interstellar medium of the host galaxy, or foreground gas along the line of sight—any of which could host turbulent

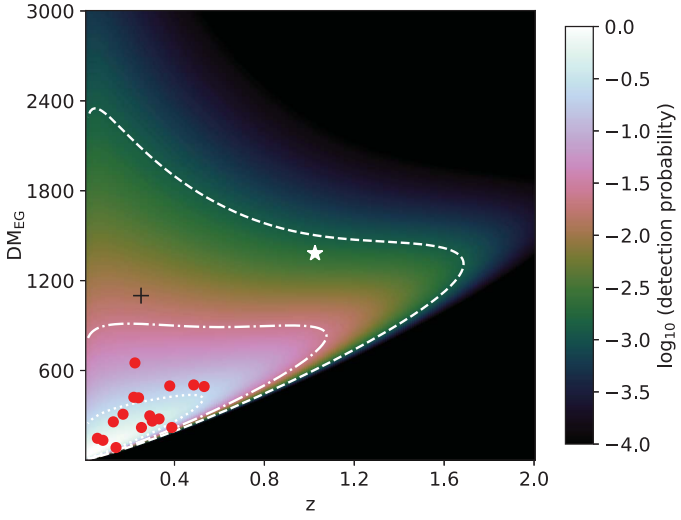


Fig. 3. Relationship between redshift and extragalactic DM for FRBs. Data are from the ASKAP incoherent sum survey (*4*). The extragalactic dispersion (DM_{EG}) is the contribution to DM after subtraction of a model for the Milky Way (*11*). Red circles indicate localized FRBs at the host galaxy redshifts. The color scale indicates an estimated detection probability, assuming a maximum energy density of $\log_{10}\left|\frac{E_\nu}{\text{erg Hz}^{-1}}\right| = 32.7 \pm 0.2$ (this work). White contours enclose 50% (dotted), 90% (dash-dot), and 99% (dashed) of the probability. The white star is FRB 20220610A. The black cross shows the repeating FRB 20190520B, which was excluded from the population inference because of differing selection effects between instruments (*28*). The uncertainties for all measurements are smaller than the symbol sizes.

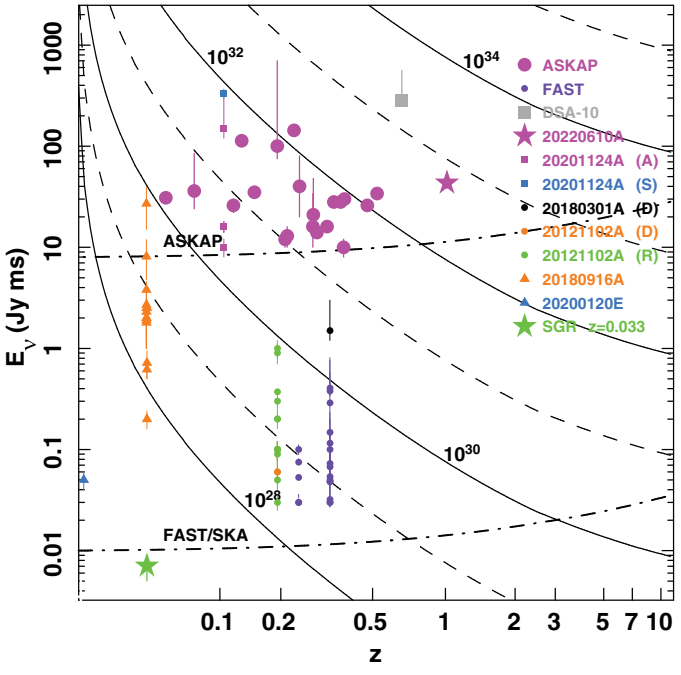


Fig. 4. Logarithmic plot of fluence as a function of redshift for localized FRBs. The magenta star is FRB 20220610A. Other notable FRBs are labeled using their Transient Name Server designations, omitting the FRB prefix for brevity. The curved solid and dashed contours indicate the energy density in units of ergs per hertz. The dash-dotted lines show the detection sensitivity of the ASKAP incoherent sum FRB search system, the Five-hundred-meter Aperture Spherical Telescope (FAST), and that expected for the Square Kilometre Array (SKA). Repeating FRBs are shown at multiple points at the same redshift but different fluences. Plotting symbols are indicated in the legend, which also indicates the FRB localization instruments. DSA, Deep Synoptic Array. The green star shows the expected fluence of the FRB-like burst from the Galactic magnetar SGR 1935+21 (*31*, *32*) if it were placed at the distance of the host galaxy of FRB 20180916A. There are biases in FRB searches that may result in the first detections of repeating sources having different properties than subsequent detections, particularly those made with other facilities (*33*). We therefore plot the repeating FRB 20201124A twice—for observations from ASKAP (designated A) (*17*) and the 25-m Stockert radio telescope (S) (*34*). For the repeating FRB 20121102A, we plot separately the initial detection (D) (*35*) and a sample of follow up bursts detected with a different instrument (R) (*36*). FRB 20180916A (*37*) and FRB 20200120E (*38*) are low-redshift repeating FRB sources. Other burst properties are reported from ASKAP (*3*, *4*, *8*, *27*), DSA-10 (*39*), and FAST observations of the repeating FRB 20180301A (*40*) and FRB 20190520D (*6*). For FRB 20180301A, we also plot the initial detection with a different instrument (D) (*41*).

magnetized plasma. For FRB 20220610A, there is an absence of strong $(\gtrsim 10^4$ rad m$^{-2})$ Faraday rotation or detectable depolarization; this is unlike the repeating FRB 20190520B, which also showed excess dispersion (*6*, *14*). We suggest that the excess dispersion for FRB 20220610A originates in a less magneto-ionically active plasma compared with that of FRB 20190520B, such as in the interstellar medium of the host galaxy rather than in the circumburst media hypothesized for other sources (*11*). Models of galaxy interstellar media imply that the DM of a typical spiral galaxy is unlikely to exceed a few hundred parsecs per cubic centimeter except in edge-on systems (*22*, *23*). Higher DM values can plausibly be produced by high-density clumps of gas within the host galaxy, particularly at $z \sim 1$, where galaxies have a substantially higher fraction of their baryons in gas (rather than stars) than at $z \sim 0$ (*24*). Alternatively, the dispersion could originate from structure in the foreground intergalactic medium or from additional ionized material associated with the possible galaxy merger between components (a), (b), and (c) discussed above.

Our DM analysis confirms inferences from other techniques (*25*) that the gas of the intergalactic medium is highly ionized. The detection of an FRB at $z > 1$ allows us to study the ionized plasma toward, around, and within the host galaxy. We expect a sight line to $z = 1$ to intersect the halos of several galaxies similar in mass and size to the Milky Way (*26*), which has previously been used to investigate foreground galaxy properties using FRBs (*27*).

Interpretation of the burst energy

The measured bandwidth-averaged fluence of FRB 20220610A is 45 ± 5 Jy ms (where Jy is the jansky, equal to 10^{-23} erg s^{-1} cm^{-2} Hz^{-1}), which implies an isotropic-equivalent spectral energy density of $(6.4 \pm 0.7) \times 10^{32}$ erg Hz^{-1} and a burst energy of $(6.4 \pm 0.7) \times 10^{41}$ erg (*11*). We derived the burst energy by assuming an intrinsic 1-GHz bandwidth and did not apply a redshift-dependent correction to the burst spectral energy distribution (a K-correction) (*11*). This value exceeds, by a factor of 3.5, the characteristic maximum energy E_{max} derived by previous FRB population models (*11*, *28*). It is unknown whether FRBs are emitted isotropically or only over a limited beaming angle, which would affect the inferred energetics and could vary between FRB sources (or between repeated FRBs from single sources). Assuming isotropic emission, we added FRB 20220610A to a previous sample of FRBs (*28*) and refitted the FRB burst energy distribution. We find that the best-fitting maximum energy E_{max} increases by a factor of 2.7 to $10^{41.7 \pm 0.2}$ erg (equivalently, the maximum energy density becomes $\log_{10}\left|\frac{E_\nu}{\text{erg Hz}^{-1}}\right| = 32.7 \pm 0.2$). In Fig. 4, we compare FRB 20220610A to the fluence of the brightest radio pulse observed from a galactic

magnetar—which is a factor of 10^5 less luminous than the burst observed from FRB 20220610A—and the wider sample of FRBs.

If a K-correction is applied, assuming a spectral index for FRB emission similar to that of the burst population found by ASKAP (*11*, *29*), we find that the burst energy integrated over the instrument bandwidth is $\sim 2 \times 10^{42}$ erg, which is higher than most localized FRBs (Fig. 4). This constrains emission models of FRBs because the electric field strength at the source can be estimated independently of the beaming angle. The calculation assumes no amplification by gravitational or plasma lensing (see supplementary text). In one class of models, FRB emission is produced near the surface of a neutron star. From the luminosity of the burst, $\sim 3 \times 10^{46}$ erg s^{-1} in the host galaxy's rest frame, we infer an electric field strength of $4.2 \times 10^{12} (r/10\ \mathrm{km})^{-1}$ V m^{-1} for a linearly polarized wave, where r (~10 km) is the curvature radius of the neutron star's magnetic field. At a neutron star surface, this value is a few percent of the Schwinger critical field strength, at which an electric field aligned parallel to the local magnetic field would be screened by electron-positron pair production (*30*). This would suppress the FRB rate above the Schwinger luminosity of $\sim 2 \times 10^{47}$ erg s^{-1} (*30*). In another class of models, FRBs are produced in a shock driven by relativistic ejecta, associated with the flare of a highly magnetized neutron star interacting with the neutron star wind. In these models, the radiative efficiency in the shock is very low ($\lesssim 10^{-5}$), and hence the required energy in the ejecta would be $\gtrsim 10^{47}$ erg, with the total flare energy being even higher. FRB 20220610A and other high-luminosity FRBs are challenging to explain in both model classes.

REFERENCES AND NOTES

1. D. R. Lorimer, M. Bailes, M. A. McLaughlin, D. J. Narkevic, F. Crawford, *Science* **318**, 777–780 (2007).
2. E. Petroff, J. W. T. Hessels, D. R. Lorimer, *Astron. Astrophys. Rev.* **30**, 2 (2022).
3. J. P. Macquart *et al*., *Nature* **581**, 391–395 (2020).
4. C. W. James *et al*., *Mon. Not. R. Astron. Soc.* **516**, 4862–4881 (2022).
5. S. Bhandari *et al*., *Mon. Not. R. Astron. Soc.* **475**, 1427–1446 (2018).
6. C. H. Niu *et al*., *Nature* **606**, 873–877 (2022).
7. R. M. Shannon *et al*., *Nature* **562**, 386–390 (2018).
8. K. W. Bannister *et al*., *Science* **365**, 565–570 (2019).
9. A. W. Hotan *et al*., *Publ. Astron. Soc. Aust.* **38**, e009 (2021).
10. C. K. Day *et al*., *Publ. Astron. Soc. Aust.* **38**, e050 (2021).
11. Materials and methods are available as supplementary materials.
12. P. A. G. Scheuer, *Nature* **218**, 920–922 (1968).
13. A. G. Mannings *et al*., Fast Radio Bursts as Probes of Magnetic Fields in Galaxies at $z < 0.5$. arXiv:2209.15113 [astro-ph.GA] (2022).
14. Y. Feng *et al*., *Science* **375**, 1266–1270 (2022).
15. D. McConnell *et al*., *Publ. Astron. Soc. Aust.* **37**, e048 (2020).
16. H. Cho *et al*., *Astrophys. J. Lett.* **891**, L38 (2020).
17. P. Kumar *et al*., *Mon. Not. R. Astron. Soc.* **512**, 3400–3413 (2022).
18. K. Aggarwal *et al*., *Astrophys. J.* **911**, 95 (2021).
19. S. Bhandari *et al*., *Astrophys. J. Lett.* **901**, L20 (2020).
20. S. Bhandari *et al*., *Astron. J.* **163**, 69 (2022).
21. Y. Guo *et al*., *Astrophys. J.* **800**, 39 (2015).
22. J. Xu, J. L. Han, *Res. Astron. Astrophys.* **15**, 1629–1638 (2015).
23. J.-F. Mo, W. Zhu, Y. Wang, L. Tang, L.-L. Feng, *Mon. Not. R. Astron. Soc.* **518**, 539–561 (2023).
24. A. Chowdhury, N. Kanekar, J. N. Chengalur, *Astrophys. J. Lett.* **941**, L6 (2022).
25. T. S. Kim *et al*., *Mon. Not. R. Astron. Soc.* **501**, 5811–5833 (2021).
26. J. X. Prochaska, Y. Zheng, *Mon. Not. R. Astron. Soc.* **485**, 648–665 (2019).
27. J. X. Prochaska *et al*., *Science* **366**, 231–234 (2019).
28. C. W. James *et al*., *Mon. Not. R. Astron. Soc.* **509**, 4775–4802 (2022).
29. J. P. Macquart *et al*., *Astrophys. J. Lett.* **872**, L19 (2019).
30. W. Lu, P. Kumar, *Mon. Not. R. Astron. Soc.* **483**, L93–L97 (2019).
31. CHIME/FRB Collaboration, *Nature* **587**, 54–58 (2020).
32. C. D. Bochenek *et al*., *Nature* **587**, 59–62 (2020).
33. P. Kumar *et al*., *Astrophys. J. Lett.* **887**, L30 (2019).
34. W. Herrmann, *The Astronomer's Telegram* **14556**, 1 (2021).
35. L. G. Spitler *et al*., *Astrophys. J.* **790**, 101 (2014).
36. J. W. T. Hessels *et al*., *Astrophys. J. Lett.* **876**, L23 (2019).
37. B. Marcote *et al*., *Nature* **577**, 190–194 (2020).
38. K. Nimmo *et al*., *Mon. Not. R. Astron. Soc.* **520**, 2281–2305 (2023).
39. V. Ravi *et al*., *Nature* **572**, 352–354 (2019).
40. R. Luo *et al*., *Nature* **586**, 693–696 (2020).
41. D. C. Price *et al*., *Mon. Not. R. Astron. Soc.* **486**, 3636–3646 (2019).
42. R. Shannon, Dataset: FRB 20220610A: Fast radio burst and host-galaxy observations, gSTAR Data Management and Collaboration Platform (2023); http://dx.doi.org/10.26185/64f7e0a9170f4.

ACKNOWLEDGMENTS

We thank K. Heintz for contributions to the FURBY project and N. Förster-Schreiber and K. Glazebrook for discussions. We thank ATNF staff for supporting the CRAFT observations with ASKAP. This work is based on observations collected at the European Southern Observatory (ESO) under programs 0105.A-0687 and 1108.A-0027. This work uses data obtained from Inyarrimanha Ilgari Bundara, the Murchison Radio-astronomy Observatory. We acknowledge the Wajarri Yamaji People as the Traditional Owners and native title holders of the observatory site. CSIRO's ASKAP radio telescope is part of the Australia Telescope National Facility (https://ror.org/05qajvd42). Operation of ASKAP is funded by the Australian government with support from the National Collaborative Research Infrastructure Strategy. ASKAP uses the resources of the Pawsey Supercomputing Research Centre. Establishment of ASKAP, Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory, and the Pawsey Supercomputing Research Centre are initiatives of the Australian government, with support from the government of Western Australia and the Science and Industry Endowment Fund. Some of the data were obtained at the W. M. Keck Observatory, which is operated as a scientific partnership among the California Institute of Technology, the University of California, and NASA. The Observatory was made possible by the generous financial support of the W. M. Keck Foundation. We recognize and acknowledge the very important cultural role and reverence that the summit of Maunakea has always had within the indigenous Hawaiian community. We are most fortunate to have the opportunity to conduct observations from this mountain. **Funding:** S.B. is supported by a Dutch Research Council (NWO) Veni Fellowship (grant no. VI.Veni.212.058). A.T.D. and K.G. acknowledge support through Australian Research Council (ARC) Discovery Project (DP) DP200102243. A.T.D. and R.M.S. acknowledge support through ARC DP DP220102305. C.W.J. and M.G. acknowledge support through ARC DP DP210102103. R.M.S. acknowledges support through ARC Future Fellowship FT190100155. J.X.P., S.S., C.D.K., A.C.G., and N.T. acknowledge support from National Science Foundation grants AST-1911140, AST-1910471, and AST-2206490. J.X.P. is a Simons Pivot Fellow. N.T. acknowledges support through FONDECYT grant no. 11191217. **Author contributions:** S.D.R., A.C.G., L.M., N.T., J.X.P., and S.S. collected and analyzed the optical and infrared imaging and spectroscopic observations. K.W.B. developed the FRB search and localization systems with contributions from S.B. and R.M.S. A.T.D., M.G., K.G., and D.R.S. performed the FRB localization and high–time resolution processing. R.D.E. and E.M.S. provided comments on the manuscript and the interpretation of the host galaxy properties. C.W.J. and W.L. interpreted the burst energetics and implications for the FRB emission mechanism. V.A.M. developed the observatory control systems that enabled FRB searches and calibration observations. H.Q. and M.W.S. measured the burst temporal properties. S.D.R., K.W.B., A.T.D., C.W.J., and J.X.P. contributed to the drafting of the manuscript. R.M.S. led the drafting of the manuscript, coordinated the FRB searches, and measured and interpreted the burst polarimetric properties. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** The raw ASKAP observations, reduced radio images and burst data, reduced optical and infrared images, and reduced optical and infrared spectra are available from the Swinburne Data Portal (*42*). The raw VLT observations are available from the ESO archive (http://archive.eso.org/eso/eso_archive_main.html) under program IDs 0105.A-0687 and 1108.A-0027. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adf2678
Materials and Methods
Supplementary Text
Figs. S1 to S8
Tables S1 and S2
References (*43*–*108*)

Submitted 12 October 2022; accepted 4 September 2023
10.1126/science.adf2678

PHOTONICS

Determining hot-carrier transport dynamics from terahertz emission

Mohammad Taghinejad[1,2], Chenyi Xia[2], Martin Hrton[3,4], Kyu-Tae Lee[5], Andrew S. Kim[5], Qitong Li[1], Burak Guzelturk[6], Radek Kalousek[3,4], Fenghao Xu[1], Wenshan Cai[5,7], Aaron M. Lindenberg[2,8], Mark L. Brongersma[1]*

Understanding the ultrafast excitation and transport dynamics of plasmon-driven hot carriers is critical to the development of optoelectronics, photochemistry, and solar-energy harvesting. However, the ultrashort time and length scales associated with the behavior of these highly out-of-equilibrium carriers have impaired experimental verification of ab initio quantum theories. Here, we present an approach to studying plasmonic hot-carrier dynamics that analyzes the temporal waveform of coherent terahertz bursts radiated by photo-ejected hot carriers from designer nano-antennas with a broken symmetry. For ballistic carriers ejected from gold antennas, we find an ~11-femtosecond timescale composed of the plasmon lifetime and ballistic transport time. Polarization- and phase-sensitive detection of terahertz fields further grant direct access to their ballistic transport trajectory. Our approach opens explorations of ultrafast carrier dynamics in optically excited nanostructures.

In nanoscale metallic systems, hot carriers—high-energy electrons and holes with a nonthermal or quasi-thermalized population exceeding a room-temperature Fermi-Dirac distribution—can be produced through the optical excitation of plasmons, followed by a nonradiative energy transfer to individual electrons located below the Fermi energy level (E_F). At the surface of a metal, some of these energetic carriers can escape and be harvested for various applications. The conversion rate from plasmons to hot carriers is predicted to be proportional to $|\vec{k}.\vec{\mathcal{E}}_{pl}(r,t)|^2$, where $\vec{k}$ and $\vec{\mathcal{E}}_{pl}(r,t)$ represent the electron wave vector and spatiotemporal profile of the plasmon electric field, respectively (*1*–*4*). This suggests an exciting possibility of creating "synthetic-transport-junctions" that can capitalize on our extensive knowledge of plasmon resonances to attain high spatial, temporal, and directional control over the distribution and transport trajectory of ejected hot carriers. These energetic carriers provide a route to coherently drive photochemical and photophysical processes (*3*, *5*–*15*). Key to their successful implementation is knowledge of the transport behavior across a junction, which has been challenging to obtain experimentally. The complexity in existing experiments stems from the complex carrier transport across the metal–dielectric junction and the thermalization of confined nonthermal carriers inside the metal. As these processes are colocated in tiny nanostructures and evolve at similar ultrafast timescales, they are not easily separated in transient optical pump-probe spectroscopies. Moreover, most transport measurements simply record the number of liberated charges into the dielectric and cannot distinguish between ballistic carriers and those that have experienced scattering. For this reason, neither the optical detection of hot-carrier–induced intraband absorption nor the electrical detection of photocurrents has the temporal resolution or capacity to perceive the ballistic motion of hot carriers as they traverse a transport junction. As a result, the application of plasmonic hot carriers has been mostly limited to actuating incoherent functionalities (*3*, *9*, *16*–*19*). To enhance their performance and to unlock the use of plasmonic hot carriers in coherent applications such as electromagnetic sources and ultrafast optoelectronic switches, understanding the underlying nonequilibrium transport physics and its correlation with plasmon fields is critical.

[1]Geballe Laboratory for Advanced Materials, Stanford University, Stanford, CA, USA. [2]Department of Materials Science and Engineering, Stanford University, Stanford, CA, USA. [3]Institute of Physical Engineering, Faculty of Mechanical Engineering, Brno University of Technology, Czech Republic. [4]Central European Institute of Technology (CEITEC), Brno University of Technology, Czech Republic. [5]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [6]X-ray Science Division, Argonne National Laboratory, Lemont, IL, USA. [7]School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [8]Stanford Institute for Materials and Energy Sciences, SLAC National Accelerator Laboratory, Menlo Park, CA, USA.
*Corresponding author. Email: brongersma@stanford.edu

Junction design and experimental setup

The exposure of a plasmonic metal–dielectric junction to a near-infrared (NIR) laser pulse initiates interfacial electron transport to produce a subpicosecond (1 ps, 10^{-12} s) current pulse containing a broad spectrum of time-varying harmonic components. An oscillating electrical current source at an angular frequency ω creates an electromagnetic wave at the same frequency (*20*). It is thus expected that the plasmonic excitation will give rise to the emission of a terahertz (THz) burst with a broad spectral signature in the terahertz (1 THz, 10^{12} Hz) band. Here, we analyze the electric field of the radiated waves as a multipotent electromagnetic probe to gain insight into the spatiotemporal dynamics of the hot-carrier transport. We consider the transport across a metal–dielectric junction created by incorporating an array of gold (Au) nanostripes in a thin titanium dioxide (TiO_2) layer. In this study (Fig. 1), Au is chosen for its excellent plasmonic properties and ability to serve as a hot-carrier reservoir. The transparent dielectric TiO_2 is selected on the basis of its wide bandgap ($E_{BG} \approx 3.44$ eV) that precludes additional photocarrier generation by the NIR photons outside of the metal. In addition, TiO_2 has a relatively large electron effective mass ($m^* \approx 0.8\ m_0$), providing a large density of available states to allow for facile injection of excited hot carriers (*1*).

We require plasmonic elements with a broken spatial symmetry. Consider a symmetric and an asymmetrically corrugated Au nanostripe embedded in a thin layer of TiO_2 as a repeating unit cell of two types of plasmonic devices (Fig. 1, A and B). Resonant illumination of these devices with NIR laser pulses excites their plasmon oscillations, which through nonradiative decay yields a nonthermal bath of hot electrons in the Au. The quadratic dependence of the plasmon-to-electron conversion rate on the plasmon field, $|\vec{k}.\vec{\mathcal{E}}_{pl}(r,t)|^2$, suggests that the generation and ensuing transport of hot carriers should spatially mimic the geometry-defined plasmonic field distribution (*1*, *21*). For symmetric stripes, this implies that an equal number of hot carriers is liberated into TiO_2 from both edges because of the mirror symmetry in the geometry and band diagram across the TiO_2/Au/TiO_2 heterostructures. By contrast, the asymmetric corrugation breaks the geometrical mirror symmetry in favor of the edge with the Au nanotips, which display plasmon-enhanced optical fields. In each unit cell, the charge transport across the Au/TiO_2 junctions produces two counterpropagating current pulses $J(t)$ within the device plane. Owing to the short transport timescale, the duration of these interfacial current pulses is expected to be subpicosecond, capable of radiating bursts of broadband THz waves (Fig. 1C). The far electric-field waveform of THz bursts is prescribed by the way charges accelerate with $\vec{\mathcal{E}}_{THz}(t) \propto \frac{dJ(t)}{dt}$ (*22*). Evidently, the temporal shape and polarization attributes of the THz field are intimately linked to the dynamics and trajectory of the interfacial charge motion. Therefore, $\vec{\mathcal{E}}_{THz}(t)$ is a sensitive electromagnetic probe that exclusively monitors the transport behavior in the presence of other transient events that lead to the internal thermalization of hot carriers.

In the device's far-field, the superposition of the two THz bursts emerging from opposing Au/TiO_2 junctions in each unit cell determine the net radiation response of the plasmonic devices. Given the much higher frequency of the NIR excitation pulse than the radiated

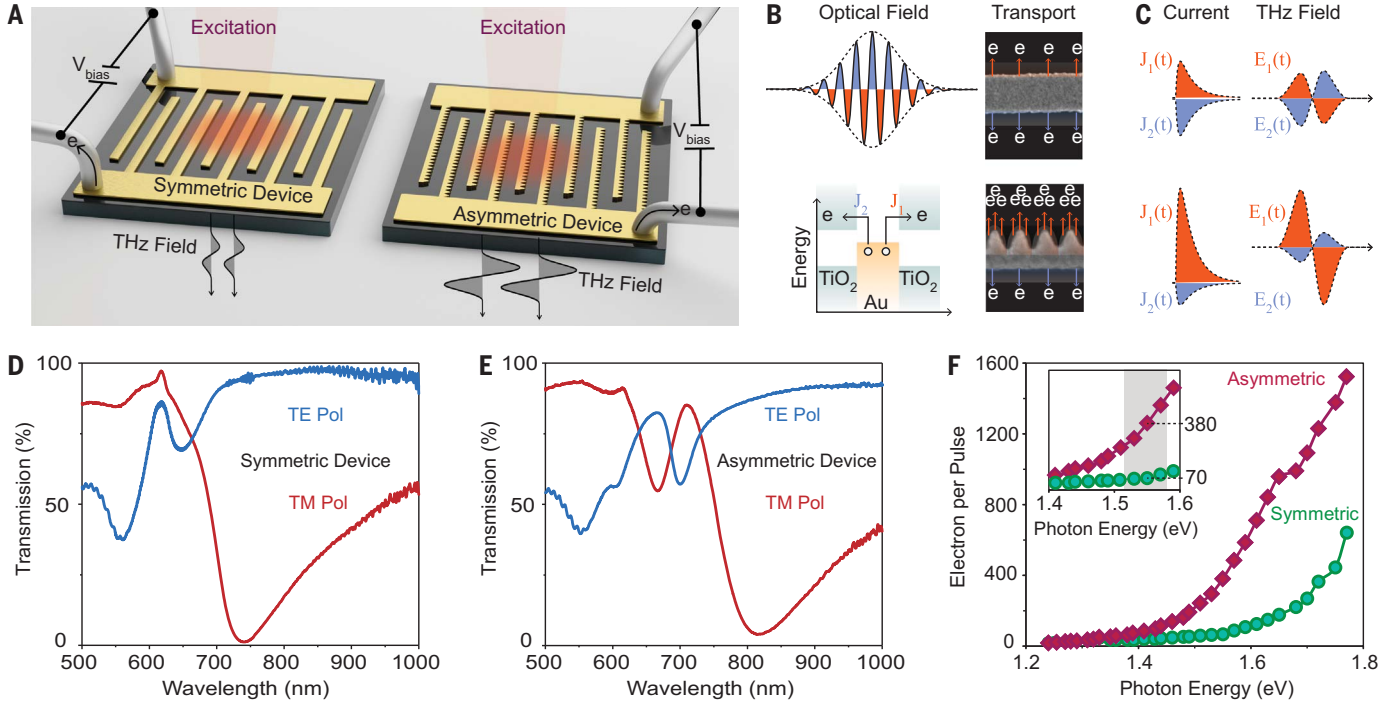


Fig. 1. Spatiotemporal interplay between plasmonic excitation field, hot-carrier transport, and radiated terahertz fields. (**A**) Schematics of two arrays of electrically enabled designer plasmonic antennas for studying the spatiotemporal dynamics of hot-electron transport through their THz radiation responses. The external bias voltages are only applied during the photocurrent measurements. (**B**) Photo-ejection of hot electrons by a NIR pulse across a TiO_2/Au/TiO_2 heterojunction. SEM images depict a symmetric and an asymmetric Au nanostripe that are embedded in a thin TiO_2 film. The symmetric structure liberates an equal number of electrons from the top and bottom junctions, enabled by the electric field of the laser in the downward (orange) and upward (blue) half cycles, respectively. Breaking the geometrical mirror symmetry leads to a higher ejection rate from the top edge owing to the enhanced field concentration at the sharp nanotips. (**C**) The interfacial transport of hot carriers produces two counterpropagating current pulses, each enabling the radiation of a coherent THz burst. The orientation and amplitude of these current pulses define the polarity and amplitude of emerged THz fields, respectively. (**D** and **E**) Measured spectra of transmitted light through the symmetric and asymmetric structures, under TE- and TM-polarized illuminations. (**F**) Number of photoemitted electrons per one laser pulse from a 1-μm-long segment of gold stripes, as a function of $\hbar\omega$. Excitation fluence, 25 μJ cm^{-2}; laser field polarization, TM. For the case of the asymmetric device, the subtle kink around 1.65 eV stems from a resonance absorption peak (fig. S3).

THz bursts, the counterpropagating current pulses on each side of a stripe evolve essentially on the same timescale. The width of the stripes is also much smaller than the wavelength of the THz emission and, consequently, the radiated terahertz fields from opposing Au/TiO_2 junctions of the symmetric structure are expected to be fully out of phase and interfere destructively (Fig. 1C). By contrast, the asymmetric concentration of the laser field on the corrugated side of the Au strips can quasi-rectify the transport direction. By breaking the geometrical symmetry, it is thus possible to generate far-field THz radiation with a polarization content that is controlled by the trajectory of the hot-carrier ejection from the gold nanotip arrays (Fig. 1C). By comparing the radiation responses of these model platforms (Fig. 1A), we can shed light on the temporal evolution and ballistic nature of the charge generation and transport.

Figure S1 shows scanning electron micrographs (SEMs) of the fabricated devices. The optical response of both devices (Fig. 1, D and E) is dominated by a plasmonic resonance around an 800-nm wavelength that is accessible through a transverse magnetic (TM) polarized illumination. This allows for a convenient plasmon excitation with a mode-locked Ti-sapphire laser in our THz spectroscopy setup. First, we study the number of hot carriers ejected into TiO_2 as a function of the excitation photon energy $\hbar\omega$. After applying a bias voltage (Fig. 1C), we illuminate the plasmonic devices with TM-polarized laser pulses (pulse width, 100 fs; repetition rate, 80 MHz; fluence, 25 μJ cm^{-2}) and record photocurrents to calculate the number of emitted hot carriers per laser pulse and by a 1-μm-long segment of the symmetric and asymmetric nanostripes (calculation details in fig. S2). In both devices, the number of ejected carriers exhibits a nonlinearly increasing trend with increasing $\hbar\omega$ (Fig. 1F) that follows from the energy dependence of the hot-carrier generation rate in Au and the photoemission rate across the Au/TiO_2 interface. According to the modified Fowler's theory, the charge injection rate over an energy barrier ϕ_B scales proportional to $(\hbar\omega - \phi_B)^2$ (*1*, *23*, *24*). In addition, the photon-to-electron conversion has a notably higher yield in the vicinity of the plasmonic resonance and intrinsic material d-band absorption. Therefore, the large density of generated hot carriers and the enhanced photoemission possibility synergistically boost the transport efficiency for increasing $\hbar\omega$. As discussed below, owing to the nature of the electronic band alignment at the Au/TiO_2 interface, these hot carriers are dominantly sp-like conduction electrons inside Au that are excited to energy states above ϕ_B upon the plasmon decay (see discussions of Fig. 2D). At the 1.55-eV photon energy, our calculation shows that one laser pulse liberates ~70 and 380 electrons from a 1-μm-long segment of symmetric and asymmetric nanostripes, respectively. This contrast, however, weakens at illumination photon energies larger than the plasmon energy because the intrinsic absorption properties of gold govern the electron-generation process.

Detecting THz emission

A schematic of our setup for the coherent detection of radiated terahertz bursts is shown in Fig. 2A (details in materials and methods). TM-polarized 800-nm laser pulses with a

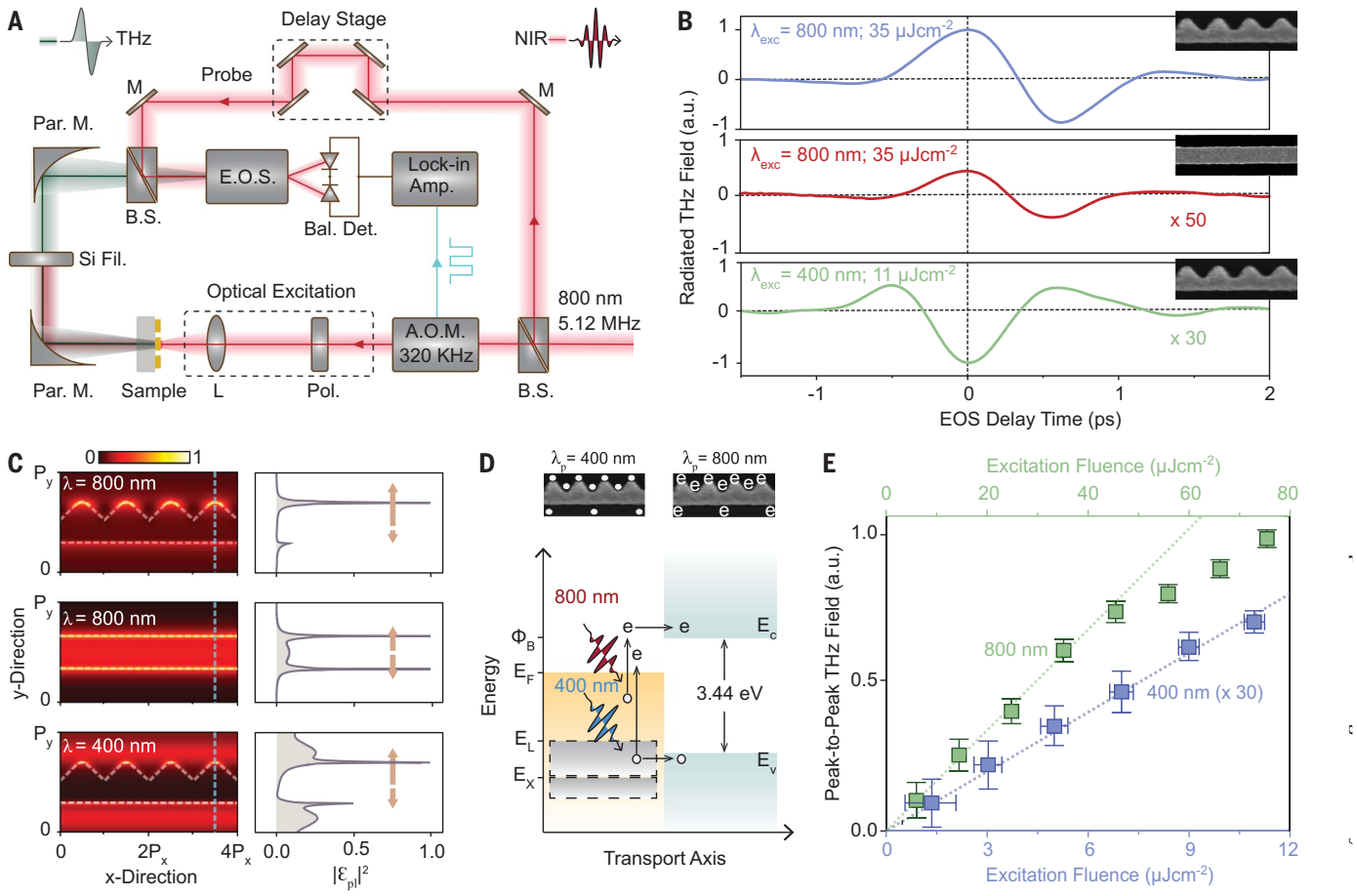


Fig. 2. Measurement of the THz radiation from rectified hot-electron and hot-hole transport. (**A**) Schematic of the free-space THz detection setup. Panels (B) and (E) are measured with a <110> cut ZnTe electro-optic crystal. (**B**) Temporal waveforms of detected THz electric fields. All waveforms are normalized to the peak field intensity in the top panel. SEM images in each inset identify the type of the characterized structure. All results presented in this figure are measured or simulated for TM-polarized illumination. (**C**) Left: Simulated spatial profiles of the electric field in the plane of Au electrodes. P_x and P_y are periodicities along *x* and *y* direction. Right: Normalized electric-field intensity (i.e., $|\mathcal{E}_{pl}|^2$) plots recorded along the dashed cyan lines on the 2D maps and normalized for a better visualization. The amplitude of current pulses propagating along the *y* direction are proportional to the intensity of the excitation laser modified by the plasmonic response. The orientation and size of the brown arrows qualitatively show the amplitude and direction of electron propagation. (**D**) A quantitative energy band diagram constructed according to a set of pump-probe spectroscopic measurements (see fig. S5 for details). E_L and E_X denote the top edges of symmetry points L and X in the d-band of gold. $\phi_B \approx 1$ eV; $E_L - E_F \approx -2$ eV; $E_X - E_F \approx -3$ eV. (**E**) Excitation fluence dependence of the peak-to-peak THz field, acquired by exciting the asymmetric device with 400- and 800-nm laser pulses. Dashed lines represent linear fits to experimental data points.

duration of $\tau_l \approx 50$ fs (fig. S4) excite plasmonic structures, and we use an electro-optical sampling (EOS) technique to detect the waveform of radiated THz bursts that copropagate with the laser beam (*22*, *25*). In this approach, $\vec{\mathcal{E}}_{THz}(t)$ is used as a gating field to induce a transient birefringence in a nonlinear EOS crystal, leading to the polarization conversion of a NIR probe from linear to elliptical. The transient polarization state of the probe is monitored to resolve the temporal waveform of the THz electric field. As illustrated in Fig. 2B, the transport of hot electrons in the asymmetric device indeed enables the radiation of a THz pulse with an electric-field waveform resembling the characteristic far-field response of a current-driven electric dipole antenna. Under a similar experimental setting, however, the detected $\vec{\mathcal{E}}_{THz}(t)$ from the symmetric structure (Fig. 2B, middle panel) is more than two orders of magnitude weaker (i.e., a 10^4 power drop). This large contrast is perhaps unexpected given that the hot-electron photoemission of the asymmetric device outperformed the symmetric structure by less than one order of magnitude. We explain this notable radiation difference as a consequence of the interplay between the spatial distribution of the optical field in the plane of the plasmonic devices and the interfacial trajectory of the hot-electron transport. A simulation of the plasmon field distribution at the 800-nm wavelength (Fig. 2C, left) indicates that the corrugated structure concentrates the incident light around the apex of the gold nanotips, whereas in the symmetric nanostripe, the top and bottom Au/TiO_2 interfaces experience an identical field enhancement. Because the spatial profile of the hot-carrier generation approximately mimics the energy density of the plasmon field, $|\vec{\mathcal{E}}_{pl}(r)|^2$ (Fig. 2C, right), the symmetric structure produces two similar current pulses that propagate in opposite directions. Given the in-plane mirror symmetry of the band diagram and device geometry, there is roughly half of one optical cycle ($T_{op} \approx 2.7$ fs) relative delay in the buildup of these current pulses that is minute in the terahertz regime. Consequently, the radiated $\vec{\mathcal{E}}_{THz}(t)$ from opposing Au/TiO_2 junctions in the symmetric device interfere destructively, as predicted above. The simulation results (Fig. 2C, right), however, suggest that the plasmon energy density at the apex of the Au nanotips is about 10 times higher than on the flat side of the corrugated nanostripe. In addition, compared to the flat Au/TiO_2 interface, the small radius of curvature of the nanotips assists to relax the

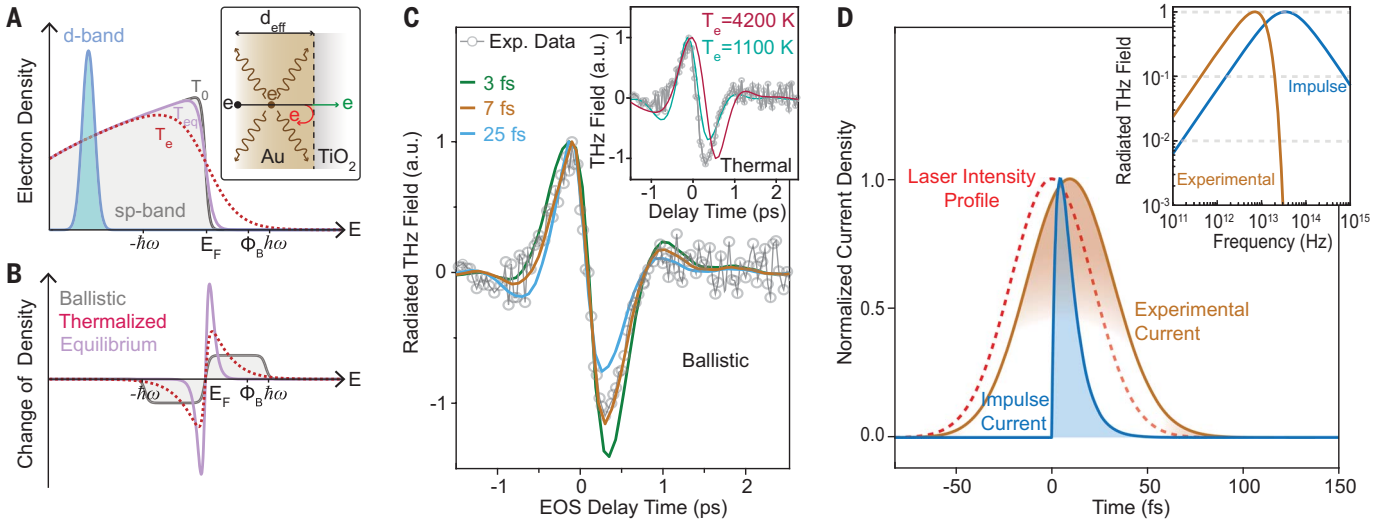


Fig. 3. Probing hot-electron transport dynamics by their THz radiation. (**A** and **B**) Qualitative illustration of electron density and its temporal evolution as a function of electron temperature upon the excitation of intraband electron transitions in gold. The decay of $\hbar\omega$ plasmons excites sp-band electrons to energy states above E_F, with a stepwise change of electron density around momentum E_F (gray). Electron-electron scattering brings the electron bath to a quasi-thermalized state that is described with a hot-electron temperature T_e (red). The transfer of energy from thermalized electrons to the Au lattice results in an equilibrium condition in the metal, described with an effective temperature, T_{eq}, larger than the room temperature, T_0. In these diagrams, E_F is assumed to be zero. (Inset) : Ballistic electrons are created in Au within an effective distance d_{eff} from the Au/TiO_2 interface. The number of ballistic electrons in this layer drops through electron-electron scattering (brown arrows), surface collision (red arrow), or ejection into TiO_2 (green arrow). (**C**) A comparison between the temporal profile of the measured THz electric field (data points) and modeled waveforms (solid lines) for three regimes of ballistic transport: τ_b = 3, 7, and 25 fs. The THz field attributed to τ_b = 3 fs is calculated when Q = 1. (Inset) Modeled waveforms produced by $J_{th}(t)$ for τ_{th} = 200 fs and T_e = 1100 and 4200 K. (**D**) Temporal shape of the hot-electron current pulses derived from the measured electric field. The extracted $J_b(t)$ under the experimental condition, τ_l = 50 fs and Q = 7, exhibits a ~58-fs pulse duration (brown). The duration of the impulse current response (blue) is ~11 fs. (Inset) Normalized amplitude of the Fourier transforms of THz waves produced by the experimental and impulse current pulses.

stringent requirement on the electron momentum matching during the transport, which furthers the photoemission efficiency (*26*). Therefore, the symmetry-breaking approach that we used produces a quasi-rectified current pulse immune from the unwanted interference effect. The weak residual radiation from the symmetric device is speculated to originate from the diminished mirror symmetry of imperfect edges and slightly oblique excitation.

Figure 2D shows an energy band diagram that is constructed with optical pump-probe spectroscopy data (see fig. S5 for details). It highlights the photon-energy considerations for transport of hot electrons and hot holes at the interface. TiO_2 has a large electronic bandgap of $E_{BG} \approx 3.44$ eV and exhibits a $\phi_B \approx 1.0$ eV barrier for the injection of electrons from Au (*27*). This barrier defines the lowest photon energy required for the photoemission of electrons through a linear process. In addition, the existence of a $E_{BG} - \phi_B \approx 2.44$ eV energy barrier for the photoemission of holes indicates the need for an even higher photon energy to liberate hot holes from Au. Therefore, the THz radiation (Fig. 2B, top and middle) produced by photons within the energy range from 1.0 to 2.44 eV primarily results from the photo-ejection of hot electrons. Such hot carriers are generated by the transition of sp-band conduction electrons from energy states below E_F of Au to excited states above ϕ_B. To study the THz radiation of the hot holes, the output of the Ti:sapphire laser is frequency doubled (see materials and methods), yielding second-harmonic photons at 400 nm with a sufficient energy to access electrons in the low-lying energy levels of Au. Indeed, 3.1-eV photons can excite interband transition of electrons residing in the d-band of Au (*28*), a process that produces hot holes with nonthermal energies up to ~0.66 eV below the valence band of TiO_2. This physical picture was verified by the observation of bursts of THz fields (Fig. 2B, bottom) after the pulsed excitation of the device with asymmetric electrodes. Similarly to electrons, hot holes also benefit from the in-plane rectification of the photoemission thanks to the asymmetric confinement of the 400-nm laser field at the gold nanotips (Fig. 2C, bottom) and the relaxation of momentum matching at the Au nanotips (*24*, *26*). Although the 400-nm laser beam is not in resonance with a plasmonic mode (fig. S3B) and the laser fluence is low, owing to a lightning rod effect at the sharp Au nanotips, the amplitude of the emerged field is notably larger than the case of the symmetric device under a strong on-resonance excitation. This observation further signifies the importance of shape engineering of the electrodes. It also highlights the valuable opportunity to capitalize on hot-hole transport in the generation and control of the radiation properties in future hot-carrier–based THz sources. This opportunity complements possible efforts to manipulate the THz emission by engineering the band structure and/or electronic states at the Au/TiO_2 interface.

Under a weak to moderate illumination, the magnitude of $\vec{\mathcal{E}}_{THz}(t)$ linearly increases with the fluence of both laser lines (Fig. 2E), implying that the driving current pulses are produced through the conversion of a single photon to a single hot carrier inside Au. However, a saturation onset for the electron transport appears as the fluence of the 800-nm beam approaches ~45 μJcm^{-2}. This effect appears to be due to the formation of a transport-induced built-in electric field (*25*) during the leading portion of the excitation pulse with a field orientation perpendicular to the Au/TiO_2 junction. The decay time of this built-in field has been reported to be a few picoseconds (*29*), hence acting as a quasi-static force resisting the photo-ejection of electrons during the tens of femtoseconds trailing portion of the excitation pulse. Such saturation effects often exist in transient phenomena that involve an interfacial transport of charges but lack a robust knob for preventing a quasi-electrostatic material charging. Nonetheless, within the available 800-nm excitation fluence in our setup, a complete saturation of the THz radiation was not observed.

Correlating THz emission and hot-carrier dynamics

From the measurements of the radiated THz field $\vec{\mathcal{E}}_{\mathrm{THz}}(t)$, we can calculate the temporal shape of the hot-electron–induced current pulse $J(t)$ with a goal of obtaining insights into the relevant timescales associated with the transport dynamics of these highly excited carriers. Such an analysis may also help to unveil the heavily debated ballistic or thermalized nature of ejected electrons (*30*). To do so, we create a physically intuitive model that captures two key transient physical processes pertinent to the buildup of $J(t)$: (i) dynamics of nonthermal hot-carrier generation under pulsed laser illumination; and (ii) energy evolution of hot electrons in Au. It is well established that ballistic nonthermal electrons are created through the excitation of surface plasmons and their subsequent nonradiative decay. The number density of generated nonthermal electrons is directly linked to the plasmon-enhanced field intensity $|\mathcal{E}_{\mathrm{pl}}(t)|^2$ inside the metal. Here, the quality factor Q of the plasmon resonance controls timescales for the buildup and decay of plasmons, both directly affecting the time span of $J(t)$. We calculate the time evolution of $\mathcal{E}_{\mathrm{pl}}(t)$ using the temporal coupled-mode theory (materials and methods) (*31*) in which Q appears in the total decay rate of the plasmon field, $\gamma_{\mathrm{d}} = \pi(QT_{\mathrm{op}})^{-1}$. Moreover, because of its highly nonequilibrium nature, the nonthermal electron system has an inherent tendency to return to its energy ground state. The equilibration occurs through scattering events that continuously alter the energy and momentum distribution of the excited electrons and consequently their ejection efficiency at the interface. Therefore, to determine the magnitude of the photo-ejected electron current $J(t)$ in our model, we need to properly account for all possible electron scattering processes that can prevent hot carriers from escaping the metallic structure, as described below.

Figure 3, A and B, depicts our phenomenological model that captures the energy evolution of hot carriers. Below E_{F} and at room temperature, Au has a relatively uniform electron density in its sp-band (Fig. 3A, gray). With a generation rate proportional to $|\mathcal{E}_{\mathrm{pl}}(t)|^2$, the nonradiative decay of $\hbar\omega$ plasmons produces a stepwise uniform distribution of nonthermal, ballistic electrons above E_{F} (Fig. 3B, gray). These ballistic charges are primarily created within an effective distance d_{eff} from the Au/TiO_2 interface (Fig. 3A, inset), half of them having an initial momentum toward the interface. Within an effective transport time $\tau_{\mathrm{t}} = d_{\mathrm{eff}} v_{\mathrm{F}}^{-1}$, where v_{F} is the electron Fermi velocity, ballistic electrons can arrive at the interface, however, with a significantly reduced number due to the emergence of electron-electron scattering in a short time frame, $\tau_{\mathrm{e-e}}$. Moreover, the severe momentum mismatch between electron wavefunctions in Au and TiO_2 causes a notable surface reflection of the ballistically transported electrons to the interface (Fig. 3A, inset) (*1*). Nonetheless, the high-energy tail of the ballistic electrons overcomes ϕ_{B} and creates a current pulse of ballistic carriers, $J_{\mathrm{b}}(t)$, over a duration $\tau_{\mathrm{b}}^{-1} = \tau_{\mathrm{e-e}}^{-1} + \tau_{\mathrm{t}}^{-1}$ as ballistic charges are removed from the metal by electron-electron scattering or when they reach the Au/TiO_2 interface. At a longer timescale, the excited electrons in the Au nanostructures return to a quasi-thermalized state in which their energy distribution follows a Fermi-Dirac statistic with an elevated electron temperature (Fig. 3A, red). This occurs after a characteristic thermalization time, τ_{th}, primarily defined by the number of electron-electron collisions needed to thermalize the electron bath. At this stage, some thermalized hot electrons with energies higher than ϕ_{b} (Fig. 3B, red) can still be injected into the TiO_2 to produce a thermal current component, $J_{\mathrm{th}}(t)$. The efficiency and dynamics of the thermal electron ejection depend on the electron temperature, which continuously evolves in time as the energy of the electron system is transferred to the lattice to finally reach a thermal equilibrium in the metal (Fig. 3, A and B, purple). Because thermalized electrons possess different energy, momentum, and spatial distributions than ballistic ones, they will display different temporal dynamics, and a different fraction of them will be emitted. We can analyze the magnitude and characteristic timescales of these dynamical processes by fitting our THz emission data to basic rate equations and a modified two-temperature model as presented in the supplementary materials (figs. S6 to S12).

To gain insight into the importance of these processes in the buildup of $J_{\mathrm{b}}(t)$, it is helpful to consider two limiting cases that have been studied in the literature (*1*). First, a $\tau_{\mathrm{b}} \approx 3$ fs case that represents the possibility of liberating ballistic hot electrons during a single optical cycle (*32*). Second, if we consider a scenario in which electron-electron scattering controls the number of escaping ballistic electrons, then a $\tau_{\mathrm{b}} \approx 25$ fs is expected given a $\tau_{\mathrm{F}} \approx 1.4$ nm fs^{-1} and an electron mean free path of 35 nm in Au (*33*). Apparently, the modeled $\vec{\mathcal{E}}_{\mathrm{THz}}(t)$ for neither of these scenarios matches the measured signal (Fig. 3C). Generally, the positive and negative half cycles of $\vec{\mathcal{E}}_{\mathrm{THz}}(t)$ are controlled by the rise and fall times of the driving current pulse, which are linked to τ_{l} and τ_{b}, respectively. As such, the observed discrepancies in the negative cycle reveal that τ_{b} is slower than subcycle photon- or field-driven photoemission processes reported in tunnel junctions (*32*) and notably faster than the characteristic timescale of electron scattering in Au. Because the plasmonic resonance displays a small $Q \approx 7$, it takes time to build up the resonance field inside the structures, and it is thus reasonable to find a value notably larger than the duration of one optical cycle but smaller than $\tau_{\mathrm{e-e}}$. Indeed, the measured THz signal is precisely reproduced using a $\tau_{\mathrm{b}} \approx 7$ fs, indicating that both the electron scattering rate and transport time inside Au influence the time span of $J_{\mathrm{b}}(t)$. By assuming literature values for $\tau_{\mathrm{e-e}}$ between 20 and 30 fs (*33*), a $\tau_{\mathrm{t}} \approx 10$ fs is calculated, suggesting that the majority of ballistic electrons are created within a $d_{\mathrm{eff}} \approx 14$ nm away from the interface. This small length scale stems from the underlying physics of the plasmon decay at the apex of Au nanotips. Previous theoretical studies predicted that in small nanoparticles, the decay of plasmons is dominated by Landau damping, a process that generates ballistic electrons within a $\tau_{\mathrm{F}} T_{\mathrm{op}}$-thick boundary layer that is ~4 nm at the 800-nm excitation wavelength (*1*). Although the extracted d_{eff} from our experiments is much smaller than the electron mean free path, it appears slightly larger than this theoretical prediction. However, because of the small radius of the curvature of Au nanotips, the portion of ballistic electrons that move away from one side of the isosceles Au triangles can still ballistically escape the metal from the other side. In addition, phonon-assisted processes may have a nonnegligible contribution to plasmon decay, which generates ballistic electrons further away from the Au/TiO_2 interface. Therefore, these two effects synergistically elongate the measured d_{eff} in our plasmonic device.

If thermalized electrons also played an important role in the charge ejection process, the THz fields would naturally show some temporal dynamics that are slower than $\tau_{\mathrm{e-e}}$. To assess this possibility, we calculated $J_{\mathrm{th}}(t)$ and its associated $\vec{\mathcal{E}}_{\mathrm{THz}}(t)$ for an experimentally estimated $\tau_{\mathrm{th}} \approx 200$ fs (figs. S9 and S10). Contrary to $J_{\mathrm{b}}(t)$, the temporal shape of $J_{\mathrm{th}}(t)$ is strongly linked to the fluence of the excitation laser. The fluence dependence of $J_{\mathrm{th}}(t)$ stems from the temperature dependence of the electron heat capacity. Indeed, as the laser fluence increases, the effective hot-electron temperature, T_{e}, increases. Considering that the electron heat capacity increases with T_{e}, at a higher laser fluence a larger amount of energy transfer to the lattice is required to lower T_{e}. Consequently, more electron-phonon collisions are needed to establish a new thermal equilibrium between electrons and the lattice that occurs over a longer timescale (*7*). To account for these effects, we modeled $\vec{\mathcal{E}}_{\mathrm{THz}}(t)$ for a wide range of electron temperatures (fig. S12) and show two drastically different cases (i.e., T_{e} = 1100 and 4200 K) in the inset of Fig. 3C. As seen, there are large discrepancies between the waveforms of the measured and calculated THz fields. Indeed, because $\tau_{\mathrm{th}} \approx 200$ fs is significantly larger than τ_{l} and electron-phonon scattering is slower than 1 ps, both positive and negative half-cycles of calculated $\vec{\mathcal{E}}_{\mathrm{THz}}(t)$ deviate from the measured signal, revealing that the rise

and decay times of the driving current pulse should be faster. In addition to their slow dynamics, the injection efficiency of thermalized electrons is much lower than that of the ballistic ones because the scattering events lower the energy of nonthermal electrons and perturb their initial momentum directed toward the Au/TiO_2 interface. In view of the described underlying physics, we suggest that the contribution of thermalized hot electrons can be safely ignored in functionalities that rely on the interfacial transport of hot carriers. The dominance of $J_b(t)$ is consistent with the linear dependence of $\vec{\mathcal{E}}_{THz}(t)$ on the laser fluence (Fig. 2E) because the number of generated ballistic carriers linearly increases with the laser fluence, whereas the thermalized counterpart is expected to display an exponential dependence (fig. S12).

From the THz waveform, we can back-calculate the temporal shape of the current pulse $J_b(t)$ and frequency content (Fig. 3D, brown curve and inset). The current pulse has a ~58-fs full width at half-maximum duration that is mostly broadened by the laser pulse. Under an impulse excitation, the calculated ballistic current pulse (Fig. 3D, blue curve) would shrink to ~11 fs, in line with the ballistic transport timescale, electron-electron scattering time, and the plasmon lifetime. The latter one defines a new opportunity for tuning the duration of the THz current pulse (and thus the radiation bandwidth) through plasmonic-mode engineering. In addition, the fast Fourier transform of the radiated field by the 58-fs current pulse exposes the prospect of achieving THz radiation above 10 THz (brown curve, inset of Fig. 3D), a bandwidth that could be expanded to above 60 THz (blue curve, inset of Fig. 3D) by using attosecond laser technology. Such an ultrabroad radiation bandwidth is currently out of reach for conventional photoconductive THz emitters because of the slow recombination lifetime of electron–hole pairs in direct bandgap semiconductors (*34*). It is also difficult to achieve through the optical rectification of short laser pulses because of the inherently mismatched group velocity of electromagnetic waves in the optical and THz regimes (*35*).

Owing to the dominance of the ballistic current component, the polarization properties of $\vec{\mathcal{E}}_{THz}(t)$ contain information about the ballistic transport trajectory of electrons and its correlation with the plasmonic field. We expose this distinctive opportunity by controlling the relative angle θ between the polarization of the excitation laser and the orientation of the asymmetric device in two sets of THz radiation experiments: rotating the laser polarization for a fixed device orientation (Fig. 4A, top) and vice versa (Fig. 4A, bottom). As seen in Fig. 4B, when the polarization angle evolves from 0° to 90° [i.e., TM to transverse electric (TE)] the peak-to-dip amplitude of $\vec{\mathcal{E}}_{THz}(t)$ drops to a minimum and then symmetrically recovers to its maximum at 180°. A similar angular pattern is measured for the hot-electron photoemission response at the Au/TiO_2 interface (Fig. 4C). These angular patterns are inherited from the dipolar nature of the plasmon resonance mode at the 800-nm resonance wavelength. A $\cos^2(\theta) + \kappa\sin^2(\theta)$ function with $\kappa \approx 0.14$ (Fig. 4B) and $\kappa \approx 0.1$ (Fig. 4C, inset) fits the angular pattern of $\vec{\mathcal{E}}_{THz}(t)$ and photocurrent, respectively, where the parameter κ is controlled by the TE over TM ratio of $|\mathcal{E}_{pl}|^2$ around the apex of Au nanotips. The translation invariance along the x axis prevents $J_x(t)$ from contributing to the far-field radiation. Therefore, this fitting function describes the interplay between the net driving current pulse perpendicular to nanostripes, $J_n(t)$, and the polarization response of the plasmon field. Upon device rotation, $\vec{\mathcal{E}}_{THz}(t)$ shows a drastically different angular behavior (Fig. 4D), particularly featuring a polarity inversion when θ exceeds ~100°. In this case, in addition to the amplitude modulation, the device rotation also steers the in-plane orientation of $J_n(t)$, denoted by the unit arrow $\boldsymbol{n}$ normal to the length of the strips in Fig. 4A. Therefore, $\vec{\mathcal{E}}_{THz}(t)$ has two orthogonal polarization components on x and y axes that are proportional to the projection of $J_n(t)$ on these axes. Because of the rectified nature of the electron transport when θ exceeds 90°, the $J_y(t)$ component experiences a flip in its direction that should appear as a π-phase shift in its associated THz field. The recorded $\vec{\mathcal{E}}_{THz}(t)$ for the rotation angles of 0° and 180° (Fig. 4D, insets) confirms this expectation because both waveforms look alike except for their inverted polarities. These physical insights guide us to a $[\cos^2(\theta) + \kappa\sin^2(\theta)][\cos(\theta) + \gamma\sin(\theta)]$ function that accurately fits experimental data when $\kappa \approx 0.14$ and $\gamma \approx 0.25$ (see calculation details in the supplementary text). The expression in the first pair of brackets simply describes the interplay between the amplitude of the total $\vec{\mathcal{E}}_{THz}(t)$ and the polarization response of

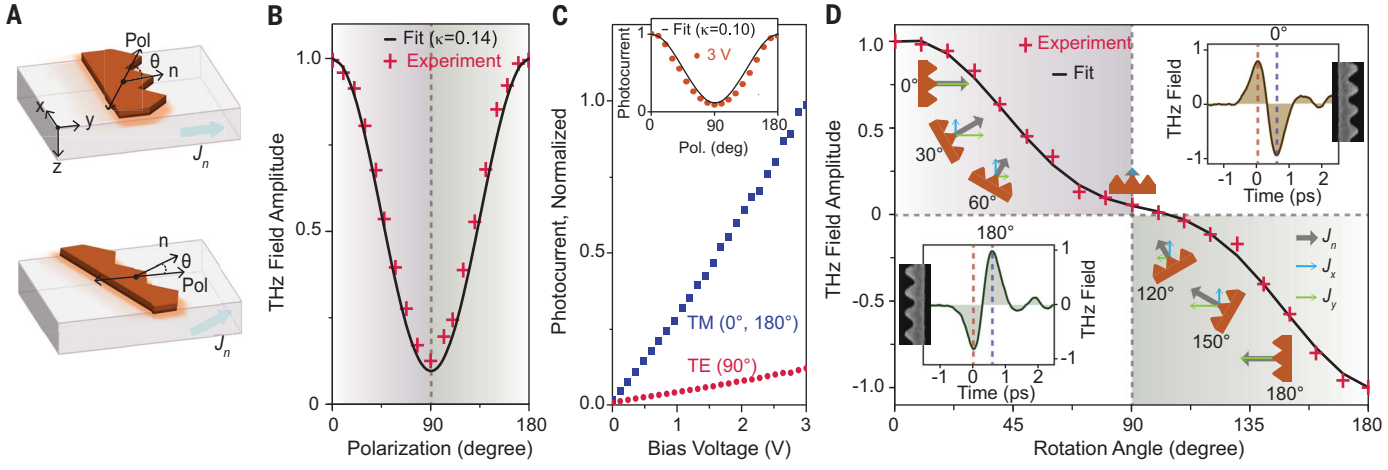


Fig. 4. Correlating the in-plane transport trajectory of hot electrons to the polarization of the illumination. (**A**) Top: Rotation of the polarization of the excitation laser under a fixed device orientation defined by the arrow $\boldsymbol{n}$, the unit vector pointing normal to the length of the electrode. Bottom: Rotation of the device orientation while the input polarization is fixed and corresponds to a TM state when θ = 0°. The green arrow denotes the direction of the current component, $J_n(t)$, produced by gold nanotips. (**B**) Terahertz field amplitude as a function of the polarization angle. Each data point is recorded by subtracting field values at 0 and ~0.6 ps, denoted by red and blue dashed lines on the insets of panel (D). (**C**) Photocurrent response of the device for the two cases of TM- and TE-polarized illuminations for an 800-nm excitation wavelength. The inset shows the angular response for a 3-V bias voltage. (**D**) Correlation between the THz field amplitude and the device orientation. Schematics show the in-plane rotation of $J_n(t)$ and its projection on the x and y axis for a few exemplary rotation angles. The THz waveforms depicted in the insets show the measured electric fields for 0° and 180° device orientations. Trailing oscillations at long electro-optical sampling times stem from phonon oscillations of environmental molecules and/or the ZnTe nonlinear crystal. Solid lines in (B) and (D) are fits according to described models.

$|\mathcal{E}_{pl}|^2$, which are linked by the angular response of $J_n(t)$. The expression in the second pair of brackets is the summation of detected x and y components of $\vec{\mathcal{E}}_{THz}(t)$, which is proportional to the summation of $J_x(t)$ and $J_y(t)$, except for a γ term that accounts for the polarization dependency of the EOS detector (see supplementary text for details). This observation demonstrates the opportunity to use $\vec{\mathcal{E}}_{THz}(t)$ as an advanced electromagnetic probe for monitoring the ballistic motion of hot electrons across transport junctions.

Concluding remarks

From the perspective of implementing a source, a hot-carrier–based THz radiator offers numerous opportunities. TiO_2 is transparent to visible and infrared photons, yet the photoemission of hot carriers facilitates the radiation of THz waves. In the presented scenario, the current pulse generation is no longer limited to photon energies larger than the electronic bandgap of the active dielectric material, as is the case for a photoconductive switch, but instead is controlled by the interfacial energy barrier. The $\phi_B \approx 1$ eV at the Au/TiO_2 interface indeed allows for the conversion of visible to NIR (~1.2 μm) photons to THz waves, owing to the fast transport of hot carriers. This wavelength range can be conveniently expanded to the short or even mid-wave infrared regime by replacing TiO_2 with, for example, silicon or indium tin oxide with smaller transport barriers (*9*, *36*). Although the magnitude of the hot-electron current pulse is smaller than the photocurrent produced in a conventional photoconductive THz source, the highly accelerated electron transport over the barrier allows for an efficient emission across a broad frequency range without applying an electrical bias, in an ultracompact device platform. Furthermore, the correlation between the near-field plasmonic response and the hot-electron transport trajectory offers new degrees of freedom for synthesizing complex polarization states, at the THz band, by relying on the rich portfolio of plasmonic platforms. The insights obtained about the spatiotemporal transport dynamics uncovers the promising potential of hot-carrier current pulses for enabling coherent optoelectronic functionalities in small-scale device architectures.

REFERENCES AND NOTES

1. J. B. Khurgin, *Nanophotonics* **9**, 453–471 (2020).
2. A. O. Govorov, H. Zhang, H. V. Demir, Y. K. Gun'ko, *Nano Today* **9**, 85–101 (2014).
3. M. L. Brongersma, N. J. Halas, P. Nordlander, *Nat. Nanotechnol.* **10**, 25–34 (2015).
4. J. B. Khurgin, *Faraday Discuss.* **214**, 35–58 (2019).
5. M. Samizadeh Nikoo *et al*., *Nature* **579**, 534–539 (2020).
6. M. Grätzel, *Nature* **414**, 338–344 (2001).
7. M. Taghinejad, W. Cai, *ACS Photonics* **6**, 1082–1093 (2019).
8. A. Polman, H. A. Atwater, *Nat. Mater.* **11**, 174–177 (2012).
9. C. Clavero, *Nat. Photonics* **8**, 95–103 (2014).
10. C. Zhu, R. E. Usiskin, Y. Yu, J. Maier, *Science* **358**, eaao2808 (2017).
11. O. Khaselev, J. A. Turner, *Science* **280**, 425–427 (1998).
12. M. Graetzel, R. A. J. Janssen, D. B. Mitzi, E. H. Sargent, *Nature* **488**, 304–312 (2012).
13. Y. Yuan *et al*., *Science* **378**, 889–893 (2022).
14. P. Christopher, H. Xin, S. Linic, *Nat. Chem.* **3**, 467–472 (2011).
15. H. Reddy *et al*., *Science* **369**, 423–426 (2020).
16. M. Taghinejad *et al*., *Nano Lett.* **18**, 5544–5551 (2018).
17. M. W. Knight, H. Sobhani, P. Nordlander, N. J. Halas, *Science* **332**, 702–704 (2011).
18. W. Li, J. Valentine, *Nano Lett.* **14**, 3510–3514 (2014).
19. G. Tagliabue *et al*., *Nat. Mater.* **19**, 1312–1318 (2020).
20. J. D. Jackson, *Classical Electrodynamics* (Wiley, ed. 3, 1998).
21. M. E. Sykes *et al*., *Nat. Commun.* **8**, 986 (2017).
22. J. Shan, T. F. Heinz, in *Ultrafast Dynamical Processes in Semiconductors* (Springer, 2004), pp. 1–56.
23. A. O. Govorov, H. Zhang, Y. K. Gun'Ko, *J. Phys. Chem. C* **117**, 16616–16631 (2013).
24. H. Chalabi, D. Schoen, M. L. Brongersma, *Nano Lett.* **14**, 1374–1380 (2014).
25. E. Y. Ma *et al*., *Sci. Adv.* **5**, eaau0073 (2019).
26. M. Grajower, U. Levy, J. B. Khurgin, *ACS Photonics* **5**, 4030–4036 (2018).
27. B. Y. Zheng *et al*., *Nat. Commun.* **6**, 7797 (2015).
28. N. E. Christensen, B. O. Seraphin, *Phys. Rev. B* **4**, 3321–3344 (1971).
29. M. Taghinejad, Z. Xu, K. T. Lee, T. Lian, W. Cai, *Phys. Rev. Lett.* **124**, 013901 (2020).
30. Y. Dubi, Y. Sivan, *Light Sci. Appl.* **8**, 89 (2019).
31. S. Fan, W. Suh, J. D. Joannopoulos, *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **20**, 569–572 (2003).
32. M. Garg, K. Kern, *Science* **367**, 411–415 (2020).
33. J. Pettine, P. Maioli, F. Vallée, N. Del Fatti, D. J. Nesbitt, *ACS Nano* **17**, 10721–10732 (2023).
34. R. A. Lewis, *J. Phys. D Appl. Phys.* **47**, 374001 (2014).
35. J. A. Fülöp, S. Tzortzakis, T. Kampfrath, *Adv. Opt. Mater.* **8**, 1900681 (2020).
36. M. Taghinejad *et al*., *Adv. Mater.* **30**, 1704915 (2018).

ACKNOWLEDGMENTS

Funding: M.T., Q.L., and M.L.B. acknowledge support from the Department of Energy (grant DE-FG07-ER46426). W.C., K.L., and A.S.K. acknowledge support from the National Science Foundation (DMR-2004749) and the Office of Naval Research (N00014-17-1-2555). C.X. and A.M.L. acknowledge support from the Department of Energy, Office of Basic Energy Sciences, Division of Materials Sciences and Engineering, under contract DE-AC02-76SF00515.
Author contributions: M.T. and M.L.B. conceived the research idea and planned the project. M.T., M.L.B., W.C., and A.M.L. devised the set of THz, optical, and optoelectronic measurements. M.T. designed plasmonic structures, performed the FDTD simulations, fabricated the devices, and performed the linear optical measurements. M.T. and C.X. performed the THz radiation spectroscopies. K.L. conducted the photocurrent measurements. A.S.K. and K.L. carried out the optical pump-probe spectroscopies. B.G. measured the transfer function of the THz spectroscopy setup. M.H. and R.K. built the analytical model for calculating the frequency response of the device. M.T., M.L.B., and A.M.L. developed the transport model. Q.L. helped M.T. with the temporal-coupled mode theory. M.T. and M.L.B. analyzed the experimental data with help from A.M.L. M.T. and M.L.B. wrote the manuscript with input from all authors. M.L.B. supervised the project.
Competing interests: The authors declare no competing interests.
Data and materials availability: All data are available in the manuscript or in the supplementary materials. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.sciencemag.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adj5612
Materials and Methods
Supplementary Text
Figs. S1 to S12
References (*37*–*42*)

Submitted 4 July 2023; accepted 4 September 2023
10.1126/science.adj5612

MARINE HEATWAVES

The collapse of eastern Bering Sea snow crab

Cody S. Szuwalski[1]*, Kerim Aydin[1], Erin J. Fedewa[2], Brian Garber-Yonts[1], Michael A. Litzow[2]

The snow crab is an iconic species in the Bering Sea that supports an economically important fishery and undergoes extensive monitoring and management. Since 2018, more than 10 billion snow crab have disappeared from the eastern Bering Sea, and the population collapsed to historical lows in 2021. We link this collapse to a marine heatwave in the eastern Bering Sea during 2018 and 2019. Calculated caloric requirements, reduced spatial distribution, and observed body conditions suggest that starvation played a role in the collapse. The mortality event appears to be one of the largest reported losses of motile marine macrofauna to marine heatwaves globally.

Snow crab (*Chionoecetes opilio*) is one of the most abundant and widely distributed species in the benthic ecosystem of the eastern Bering Sea. The snow crab population has supported an iconic fishery valued at ~US$150 million (ex-vessel average from 2012 to 2021), and fisheries management in Alaska is considered some of the most effective in the world (*1*). A key component of this management is a yearly bottom trawl survey conducted by the National Marine Fisheries Service that monitors the size and number of snow crab on the eastern Bering shelf (Fig. 1A). Many field and laboratory studies aimed at understanding population processes such as growth and maturity have also been performed (*2*). Despite this attention and effort, the stock unexpectedly collapsed in 2021.

The collapse in 2021 occurred 3 years after the observed abundance of snow crab was at historical highs (Fig. 1C). Groups of crab of similar sizes are called "pseudocohorts" because true cohorts cannot be identified as a result of difficulties in aging crab associated with the loss of the hard body parts during the molting process. The largest pseudocohort on record began to be observed in the survey beginning in 2015 and unexpectedly declined by roughly half from 2018 to 2019 (Fig. 1D). The survey was cancelled in 2020 because of the coronavirus pandemic. The 2021 survey found the fewest snow crab on the eastern Bering Sea shelf since the survey began in 1975. More than 10 billion crab disappeared from the eastern Bering Sea shelf from 2018 to 2021 (*3*).

Hypotheses to explain the disappearance fall under two categories: Either the crab are still alive, but the survey did not sample them, or the crab died. It is possible that the crab are in the eastern Bering Sea but were poorly sampled by the most recent surveys. If this were the case, one would expect estimates for other morphometrically similar species such as Tanner crab to have declined unexpectedly, but the population trend for Tanner crab increased (fig. S1). Movement to the northern Bering Sea could account for declines in the eastern Bering Sea, but surveys in the northern Bering Sea did not find crab in the quantities or of the correct sizes to explain declines in the southeast (Fig. 1A). Movement west into Russian waters is another possibility, but Russian scientists reported declines in catch per unit effort in 2020 (*4*), which one might not expect if crab from Alaska emigrated. In addition, it is possible that the crab moved into deeper waters on the Bering Sea slope. The high fishery catch per unit effort in deeper waters during the 2021 fishery supports this possibility to some extent, but the amount of available habitat is <10% of that on the shelf (*5*), and fishery catch per unit effort during 2022 was the lowest on record (fig. S2). Consequently, it is unlikely that all of the missing crab from the shelf are on the slope. Given these observations, mortality is a likely culprit for the bulk of the collapse.

Changes in temperature, predation, fishery effects, disease, and/or cannibalism could affect mortality rates. Snow crab are generally associated with cold water, but they can function in waters up to 12°C in the laboratory (*6*). A marine heatwave occurred in the Bering Sea during 2018 and 2019, and the "cold pool" (a mass of water <2°C on the sea floor with which juvenile snow crab are associated) was absent during this period (Fig. 1B). Although not fatal, the resulting bottom temperatures could affect metabolic rates and alter intra- and interspecific interactions. Smaller crab are a major component in the diet of Pacific cod in the Bering Sea (*7*), and recent changes in the distribution and abundance of cod and crab have contributed to increased consumption of crab by cod (fig. S3). Removals by the snow crab fishery and incidental mortality in fisheries for other species in the Bering Sea may also affect the population dynamics of snow crab. Cannibalism of smaller crab by larger crab is another possible source of mortality and has been suggested as an important driver of population dynamic[illegible] eastern Canadian populations (*8*). Last, [illegible]ter crab syndrome, a fatal disease that results from infection by a dinoflagellate (*9*), has been observed more frequently in the survey in the past several years and is generally associated with warmer conditions and high densities of immature crab.

To understand the recent collapse, we first attempted to understand the historical variability in mortality. We fitted a population dynamics model to the abundance and size composition data for male crab and estimated recruitment (small crab entering the population) and a maturity- and year-specific total mortality. "Total mortality" represents the fraction of crab dying from any cause in a given year. We then collated maturity-specific time series of potential stressors from 1990 to 2019 and used them in generalized additive models (GAMs) (*10*) to predict total mortality estimated from the population dynamics models (detailed methodology, sensitivities, and simulation testing are provided in the supplementary materials, materials and methods).

The population dynamics model fitted the indices of abundance and size composition data from the survey well, which is not unexpected, given the flexibility of the model (Fig. 2, A and B). Estimated mortality was higher and more varied for mature crab than for immature crab, and estimated mortalities in 2018 and 2019 were the some of the highest in the time series for immature and mature crab. We simulated snow crab populations with time variation in mortality to understand the ability of our population dynamics model to estimate these quantities with the available data. The correlation between estimated mortality and simulated mortality were high, which suggests that analyses relating estimates of mortality and environmental covariates are justifiable (supplementary materials, materials and methods).

GAMs fitted to estimated immature and mature mortality explained ~78 and ~71% of the variability, respectively (Fig. 2C). Higher temperatures and higher densities of mature crab were significantly associated with higher estimated mortality for mature crab. Higher temperatures were also associated with higher immature mortality, but the best-fitting relationship between immature mortality and mature density was dome shaped. The importance of temperature and density was robust to changes in data quantity (such as leave-one-out cross validation), consideration of the uncertainty in mortality estimates, and randomization trials (supplementary materials, materials and methods).

Our results do not support a strong connection between variability in snow crab mortality and indices of trawling, predation, cannibalism, or disease. All of these forces must contribute to underlying mortality to some degree,

[1]Alaska Fishery Science Center, National Oceanic and Atmospheric Administration, Seattle, WA, USA. [2]Alaska Fishery Science Center, National Oceanic and Atmospheric Administration, Kodiak, AK, USA.
*Corresponding author. Email: cody.szuwalski@noaa.gov

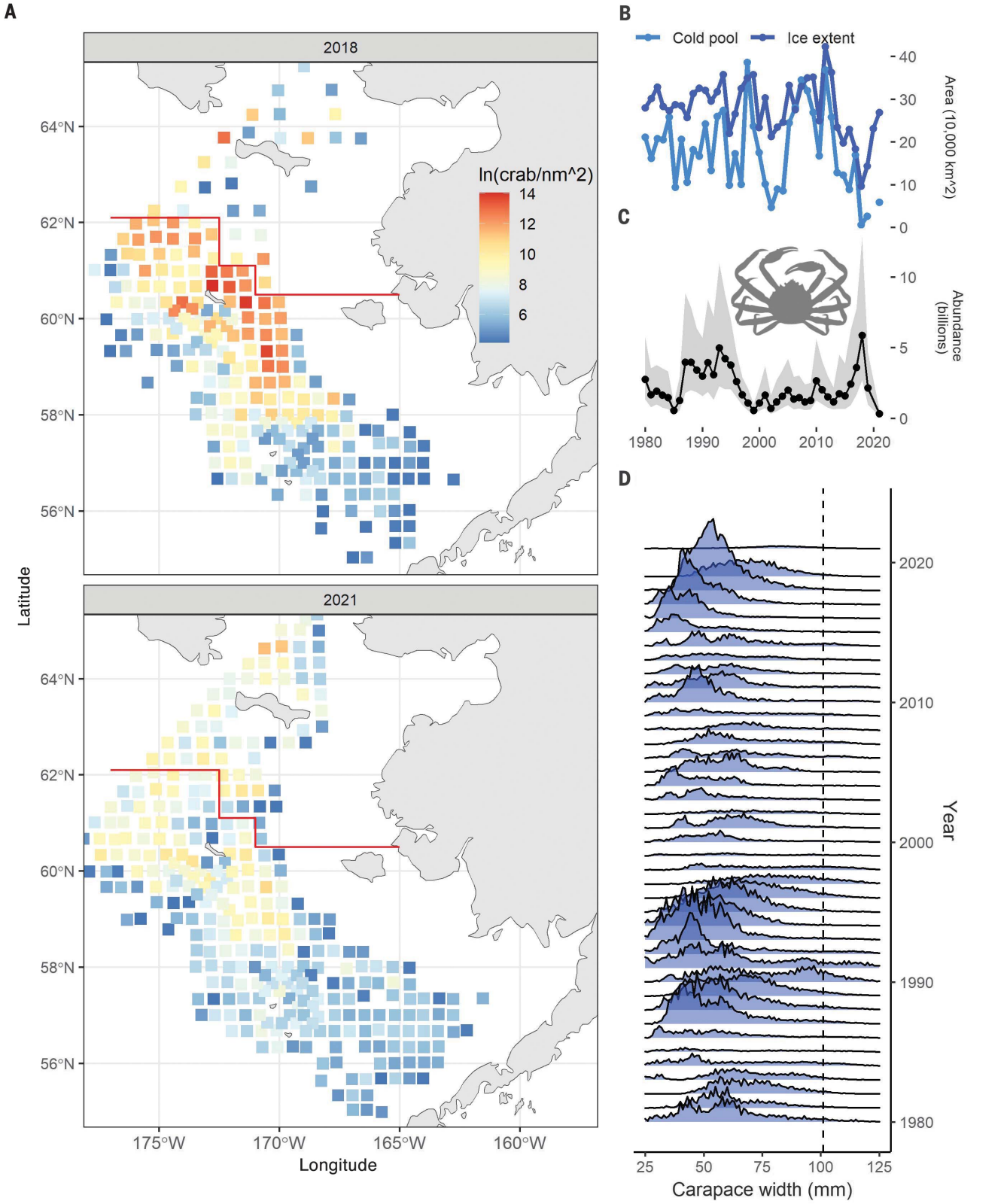


Fig. 1. The collapse of the snow crab population in the eastern Bering Sea. (**A**) Snow crab are widely distributed on the eastern Bering Sea shelf, and densities of crab were an order of magnitude lower in 2021 compared with 2018. Each square indicates a survey tow with snow crab present, and the red line separates the northern Bering Sea survey and eastern Bering Sea survey extent. lat, latitude; lon, longitude. (**B** and **C**) Changes in ice extent and (B) the resulting cold pool area influence (C) the population dynamics of snow crab. Only male abundance is plotted; shading indicates the 95% confidence intervals. (**D**) The relative numbers at size of crab observed in the survey over time. The vertical dashed line indicates the size at which the fishery begins to retain crab. Smaller crab are poorly captured by the survey gear, so the true size of a cohort is not apparent until a few years after it is first observed in the survey—compare 2015 with 2018, for example.

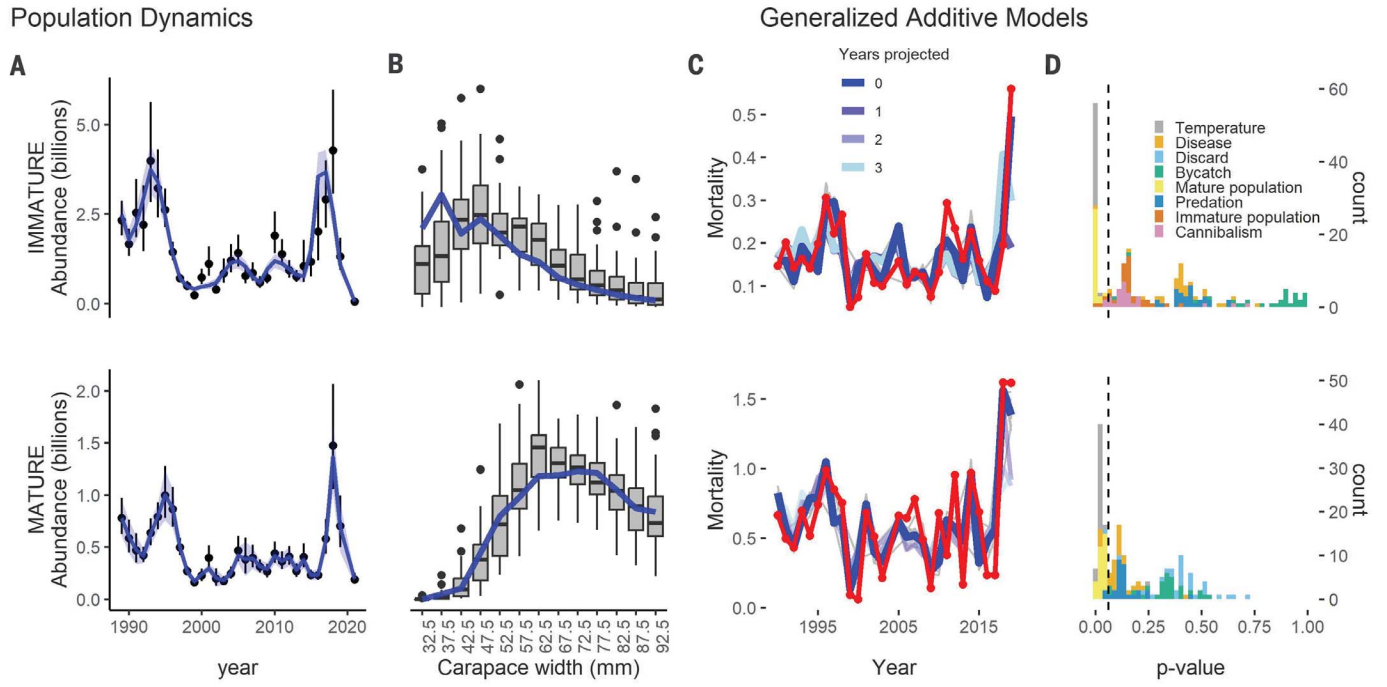


Fig. 2. Population dynamics fits and covariate relationships with mortality. Population dynamics model fits to the survey data. The top row relates to immature crab; the bottom row relates to mature crab. (**A**) Points and vertical lines indicate observed abundances, and estimates are shown in blue, with 95% confidence intervals shaded in light blue. (**B**) Box plots are size composition data aggregated over the study period with average estimated size composition over year by size shown in blue. (**C**) Fits (blue) to estimated mortality (red) from GAMs with (**D**) the significance of covariates resulting from replicates over leave-one-out cross validation. Lighter shades of blue in (C) are fits of the model with a given number of years excluded from the end of the time series and then predicting the held-out data.

but several observations corroborate the idea that temperature and population density were the key variables in the recent collapse. Cod predation was near average levels during the collapse when considering the number of snow crab in the system, and for a portion of the year, a large fraction of the cod population moved out of the eastern Bering Sea into the northern Bering Sea, which rarely occurs. This movement would serve to decrease the already average predation on snow crab in our study area. Furthermore, the missing snow crab were of a size that is larger than cod usually eat in large numbers. Trawling pressure has been another oft-raised potential cause of the collapse, but trawling activity in the Bering Sea has been relatively consistent over the past ~20 years. The observed bycatch of snow crab by trawlers was several times lower during the collapse than historical highs, and estimated mortality was low during the historical periods of high bycatch. This observation, coupled with the largest cohort in history having established and grown on the sea shelf for ~8 years under consistent trawling pressure, are difficult to reconcile with the idea that trawling contributed to the collapse. A more in-depth discussion of these points is available in the supplementary text.

Assessing the predictive skill of a model is an important check on overfitting and is relevant to providing management advice. After an ecologically and economically costly collapse, it is natural to ask whether we could have foreseen the collapse. To explore this question, we excluded 1, 2, and 3 years of data from the end of the time series; refitted the models; then tried to predict the excluded years of mortality with the covariates from those years. The model for immature mortality contained enough information in 2016 to forecast an increase in mortality, but it was not able to reach the magnitude of the estimated mortality in 2019 (Fig. 2C). The model for mature mortality performed similarly, forecasting an increase in mortality over the projection period, but it was not able to reach the estimated mortalities until the most recent data were included in the model. This suggests that the circumstances underpinning the recent collapse were unprecedented in the Bering Sea in recent history.

The collapse of the eastern Bering Sea snow crab population appears to be one of the largest reported losses of motile marine macrofauna to marine heatwaves globally (*11*), exacerbated by the record number of snow crab in the ecosystem. However, the thermal limits of snow crab exceed the observed temperatures (*6*). Temperature-dependent caloric requirements are a potential explanation to relate temperature to mortality. Foyle *et al.* (*6*) showed that caloric requirements for snow crab in the laboratory nearly double from 0°C to 3°C, which is roughly the change experienced by immature crab from 2017 to 2018 (Fig. 3A). Caloric requirements in 2018 (calculated from temperature occupied, abundance of crab at size, and weight at size) quadrupled for the modeled fraction of snow crab in the eastern Bering Sea from 2017 and were double the previous maximum value in 1989 (Fig. 3B). The impact of increased caloric demands appears to be reflected in the observed weight at size. A crab with a 75-mm carapace width in 2018 weighed on average 156 g and was ~25 g lighter (~15% of its body weight) than a crab in 2017 of the same size in the same-temperature waters (Fig. 3C). Furthermore, the spatial footprint of the stock was near the lowest levels historically in 2018 (Fig. 3, D and E). The unprecedented caloric demands, coupled with a small area from which to forage relative to historical grounds, suggest that starvation likely played a role in the disappearance of more than 10 billion snow crab, similar to the marine heatwave–related collapse of Pacific cod in the Gulf of Alaska in 2016 (*12*).

The eastern Bering Sea snow crab population collapsed once before in the late 1990s, but that collapse arose from a lack of recruitment, not a sudden mortality event. The Arctic Oscillation and sea ice have been linked to

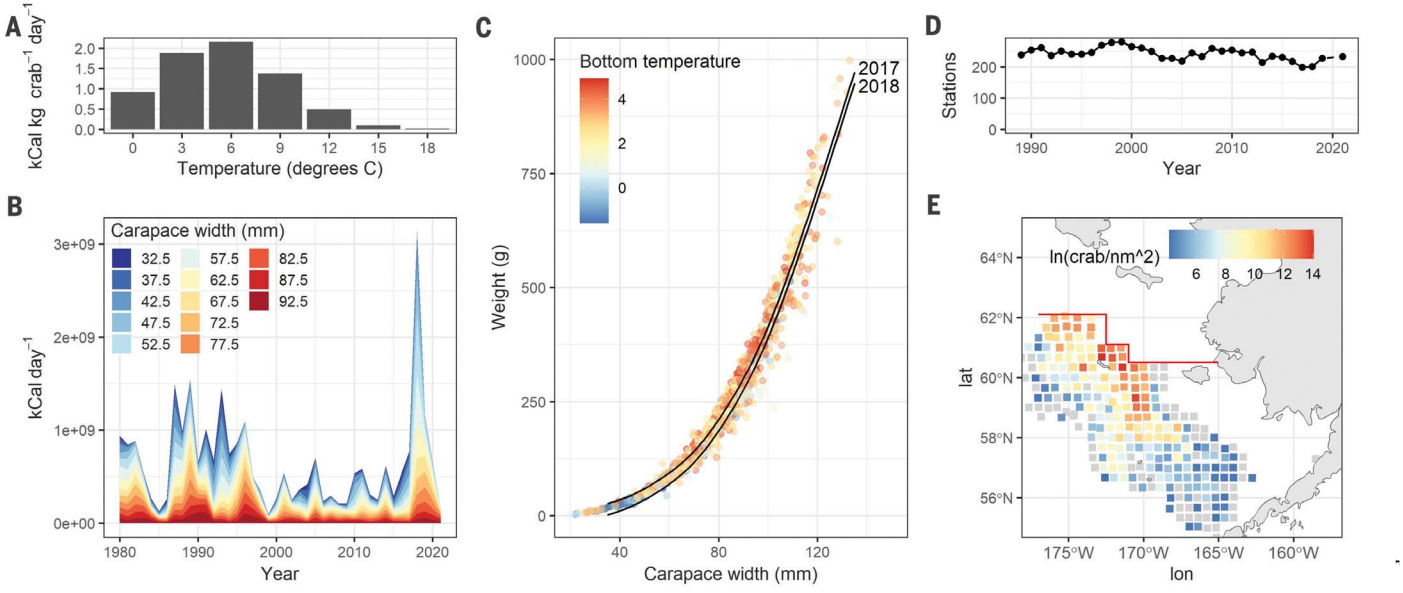


Fig. 3. Influence of temperature on snow crab metabolism, body condition, and distribution. (**A**) Impact of temperature on caloric requirements for snow crab in the laboratory. [Reproduced from Foyle *et al*. (*6*).] (**B**) The extrapolated caloric requirements for crab in the eastern Bering Sea based on temperature, abundance at size, and weight at size. (**C**) The observed weight at size colored by the temperature (°C) at which the crab were collected. The lines indicate the relationship between weight at size in 2017 and 2018 while holding temperature at 1°C. (**D**) The spatial extent of the stock has varied over time, as seen through the number of stations at which crab were observed in the 400-square-nautical-mile grid. (**E**) Distribution in 2018 (colored tiles) was one of the smallest on record (gray tiles indicate maximum historical range).

snow crab recruitment, and projections of recruitment suggest that snow crab abundances will decline in the future as sea ice disappears from the eastern Bering Sea (*13*). However, these declines were projected to occur at least 20 years from now. Given the recent collapse, the short-term future of snow crab in the eastern Bering Sea is precariously uncertain. Over the long term, the northern Bering Sea is a prospective climate refuge for snow crab (and potentially a fishery) (*14*), but the possibility of a fishery rests on the unclear prospect of crab growing to a larger size in the north and the currents retaining pelagic larvae released in the northern Bering Sea.

In 2020 and 2021, 59 boats fished for snow crab, which brought $227 million (ex-vessel) into fishing communities (*15*). The disappearance of snow crab will be a staggering blow to the functioning of some communities in rural Alaska, such as those on Saint Paul Island, which rely strongly on the revenue derived from the capture and processing of snow crab. The Magnuson-Stevens Act includes provisions for fisheries disaster assistance, which were designed to provide economic support for communities who face hardship as a result of collapsed fisheries. The number of applications in the United States has been increasing in recent years (*16*), and an application for snow crab was received in early 2022. These funds are a boon in the medium term, but years can pass between disaster and dispersal of these funds. Consequently, Alaskan crabbers face an uncertain short-term future because the disaster funds may not arrive in time to forestall the bankruptcy of long-standing businesses.

Beyond the fishery for snow crab, Alaskan fisheries are some of the most productive in the world; they produced 5.27 billion tons of seafood in 2021, valued at US$1.9 billion (*17*). The collapse of the snow crab population was a strong response to a marine heatwave, and other populations in the Bering Sea also suffered large losses. Salmon populations in the north collapsed, and seabird and seal die-offs occurred (*18*). However, other populations are flourishing. Sablefish abundances are at all-time highs in the Bering Sea (*19*), and the assessment for walleye pollock, which supports the largest fishery in the Bering Sea and one of the largest in the world (*20*), reported one of the largest estimated year classes established in 2018 (*21*). Pollock may still decline under continued warming (*22*), but the short-term response to markedly warmer bottom waters was a population boom. The adaptive capacity of species is a key uncertainty in the outcome of warming oceans, but it is very probable that the benthic community in the eastern Bering Sea in the not-too-distant future will look different from today's given the rapid pace of warming (*23*).

Overfishing has historically been the largest threat to global fisheries, but in many parts of the world, this problem has been addressed with careful management (*1*). Climate change is the next existential crisis for fisheries, and snow crab are a prime example for how quickly the outlook can change for a population. In 2018, catches were projected to increase to levels not seen in decades. Three years later, the population had collapsed. Our current management tools base management targets and projected sustainable yields on the historical dynamics of a population. However, projections based on historical dynamics are unreliable when the future environmental conditions of a region will not resemble the past. Incorporating environmental drivers into management targets has been a recent focus of the scientific literature, but this can result in counterintuitive management responses such as increasing exploitation rates on populations undergoing climate-related declines in productivity (*24*). Our experience in the management of a collapsing snow crab population suggests that considering environmental influence in estimates of biomass used to set catch limits can be important, but how to consider environmental change in management targets is an unresolved question.

Beyond reconsidering how sustainable catches and management targets are calculated under widespread changes in productivity, other practical matters need close attention from fisheries managers and stakeholders. These include (i) enacting efficient and timely disaster response with plans for the possibility that a "disaster" is permanent, (ii) implementing management institutions that allow fishers to pursue diverse

portfolios of species, and (iii) ensuring consistent and timely biological surveys. Support for the development of alternative marine-based livelihoods (such as mariculture) may also alleviate some of the pressures associated with fishery collapses. The Bering Sea is on the frontlines of climate-driven ecosystem change, and the problems currently faced in the Bering Sea foreshadow the problems that will need to be confronted globally.

REFERENCES AND NOTES

1. R. Hilborn *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **117**, 2218–2224 (2020).
2. L. A. Copeman *et al*., *Polar Biol.* **44**, 1883–1901 (2021).
3. C. S. Szuwalski, "Stock assessment of Eastern Bering Sea snow crab. Stock assessment and fishery evaluation report for the king and tanner crab fisheries of the Bering Sea and Aleutian Islands regions" (North Pacific Fishery Management Council, 2021).
4. I. S. Chernienko, *Izvestiya TINRO* **210**, 359–370 (2021).
5. G. R. Hoff, "Results of the 2016 eastern Bering Sea upper continental slope survey of groundfish and invertebrate resources," NOAA technical memorandum NMFS-AFSC-339 (2016).
6. T. P. Foyle, R. K. O'Dor, R. W. Elner, *J. Exp. Biol.* **145**, 371–393 (1989).
7. G. M. Lang, P. A. Livingston, "Food habits of key groundfish species in the eastern Bering Sea slope regions," NOAA technical memorandum NMFS-AFSC-67 (1996).
8. G. A. Lovrich, B. Sainte-Marie, *J. Exp. Mar. Biol. Ecol.* **211**, 225–245 (1997).
9. T. R. Meyers *et al*., *Dis. Aquat. Organ.* **26**, 221–227 (1996).
10. S. N. Wood, *J. R. Stat. Soc. Series B Stat. Methodol.* **73**, 3–36 (2011).
11. K. E. Smith *et al*., *Ann. Rev. Mar. Sci.* **15**, 119–145 (2023).
12. S. J. Barbeaux, K. Holsman, S. Zador, *Front. Mar. Sci.* **7**, 703 (2020).
13. C. S. Szuwalski *et al*., *ICES J. Mar. Sci.* **78**, 502–515 (2020).
14. E. J. Fedewa, T. M. Jackson, J. Richar, J. L. Gardner, M. A. Litzow, *Deep Sea Res. Part II Top. Stud. Oceanogr.* **181–182**, 104878 (2020).
15. B. Garber-Yonts, J. Lee, "SAFE report for the king and tanner crab fisheries of the Gulf of Alaska and Bering Sea/Aleutian Islands Area: Economic status of the BSAI king and tanner crab fisheries off Alaska, 2021" (North Pacific Fishery Management Council, 2021).
16. L. Bellquist, V. Saccomanno, B. X. Semmens, M. Gleason, J. Wilson, *PeerJ* **9**, e11186 (2021).
17. NOAA Fisheries, Fisheries of the United States Report (NOAA, 2022); https://www.fisheries.noaa.gov/foss.
18. E. C. Siddon, S. G. Zador, G. L. Hunt Jr., *Deep Sea Res. Part II Top. Stud. Oceanogr.* **181–182**, 104914 (2020).
19. D. Goethel *et al*., "Assessment of the sablefish stock in Alaska. NPFMC Bering Sea, Aleutian Islands and Gulf of Alaska Stock Assessment and Fishery Evaluation report" (North Pacific Fishery Management Council, 2022).
20. Food and Agriculture Organization, *The State of World Fisheries and Aquaculture 2022. Towards Blue Transformation* (Food and Agriculture Organization, 2022).
21. J. Ianelli *et al*., "Assessment of the walleye pollock stock in the eastern Bering Sea. NPFMC Bering Sea, Aleutian Islands and Gulf of Alaska Stock Assessment and Fishery Evaluation report" (North Pacific Fishery Management Council, 2022).
22. F. Mueter, N. Bond, J. Ianelli, A. Hollowed, *ICES J. Mar. Sci.* **68**, 1284–1296 (2011).
23. M. Rantanen *et al*., *Commun. Earth Environ.* **3**, 168 (2022).
24. C. S. Szuwalski *et al*., *Fish Fish.* **24**, 439–453 (2023).
25. C. S. Szuwalski, K. Aydin, E. J. Fedewa, B. Garber-Yonts, M. A. Litzow, The collapse of eastern Bering Sea snow crab Zenodo (2023); https://zenodo.org/record/8184862.

ACKNOWLEDGMENTS

We thank the Resource Assessment and Conservation Engineering Division and the Kodiak Shellfish Assessment Program, at the Alaska Fisheries Science Center, and the Alaska Department of Fish and Game for long-term data collection efforts, without which analyses such as this would be impossible. We also thank the Bering Sea Fisheries Research Foundation for selectivity data from their trawl experiments. A. Hollowed, S. Zador, L. Barnett, S. Barbeaux, the members of the North Pacific Fishery Management Council's Crab Plan Team, the Science and Statistical Committee, and three anonymous reviewers provided comments that improved this manuscript. **Author contributions:** Conceptualization: C.S.S., K.A., E.J.F., B.G.-Y., and M.A.L. Methodology: C.S.S. and K.A. Formal analysis: C.S.S. Investigation: C.S.S., K.A., E.J.F., B.G.-Y., and M.A.L. Visualization: C.S.S. Writing – original draft: C.S.S. Writing – review and editing: C.S.S., K.A., E.J.F., B.G.-Y., and M.A.L. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** A repository including the data and code used to perform the analysis can be found at (*25*). **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adf6035
Materials and Methods
Supplementary Text
Figs. S1 to S51
Tables S1 to S6
References (*26–45*)
MDAR Reproducibility Checklist

Submitted 4 November 2022; accepted 31 August 2023
10.1126/science.adf6035

POLYMER CHEMISTRY

Chemically recyclable polyolefin-like multiblock polymers

Yucheng Zhao[1]†, Emma M. Rettner[2]†, Katherine L. Harry[1], Zhitao Hu[1], Joel Miscall[3,4], Nicholas A. Rorrer[3,4], Garret M. Miyake[1,2]*

Polyolefins are the most important and largest volume plastics produced. Unfortunately, the enormous use of plastics and lack of effective disposal or recycling options have created a plastic waste catastrophe. In this work, we report an approach to create chemically recyclable polyolefin-like materials with diverse mechanical properties through the construction of multiblock polymers from hard and soft oligomeric building blocks synthesized with ruthenium-mediated ring-opening metathesis polymerization of cyclooctenes. The multiblock polymers exhibit broad mechanical properties, spanning elastomers to plastomers to thermoplastics, while integrating a high melting transition temperature (T_m) and low glass transition temperature (T_g), making them suitable for use across diverse applications (T_m as high as 128°C and T_g as low as −60°C). After use, the different plastics can be combined and efficiently deconstructed back to the fundamental hard and soft building blocks for separation and repolymerization to realize a closed-loop recycling process.

Polyolefins make up over half of all plastics produced with diverse applications that impact nearly every aspect of modern life (*1*, *2*). However, the inert characteristics that provide plastics their useful materials properties make them resilient to degradation in the environment, with an estimated lifetime of hundreds of years and unknown environmental consequences (*3*, *4*). Mechanical recycling of plastics is largely unsuccessful for preserving their value or as a waste remediation strategy; only around 9% of all plastics have been recycled (*5*). A substantial challenge in mechanical recycling is the need for physical separation because mixing different types of plastics leads to macrophase-segregated materials with compromised materials properties (*6*). Even if separation is achieved, mechanical recycling initiates cross-linking and chain-scission reactions that abate materials properties. Valorization of plastics is an alternative approach to recapture resources but yields new chemicals rather than original monomer feedstocks (*7*, *8*). For example, polyethylene (PE) can be converted to low–molecular weight hydrocarbons for new applications including liquid fuels or waxes (*9*–*11*). An inspirational approach for a closed-loop life cycle of sustainable plastics is to design systems that are chemically depolymerizable back to the original monomers for purification and repolymerization (*12*). These technologies are not yet economically or technologically viable replacements for polyolefins because the complete depolymerization back to monomer maximizes the number of chemical bonds that must be broken and reformed during the recycling process, necessitating extreme reaction efficiency and increased energy inputs (*13*, *14*). A more efficient strategy would deconstruct plastics back to oligomeric units to allow for purification, yet minimize the number of bond transformations that must occur during chemical recycling.

Despite strong motivations, depolymerizable polyolefin-like materials have not yet been established as a broad replacement for the rich spectrum of polyolefin applications (*15*). The deconstruction of polyolefins is thermodynamically challenging, requiring energy-intensive conditions to break strong carbon-carbon bonds, leading to substantial side reactivity. As such, the incorporation of more chemically reactive functional groups into the polyolefin main chain to facilitate degradation has gained increasing attention (*16*, *17*). Insertion of ketones into PE through the copolymerization of ethylene with carbon monoxide results in photodegradable PE derivatives that can break down to smaller fragments but falls short of creating a circular plastics life cycle (*18*). The step-growth condensation polymerization of diester-functionalized alkanes with diols results in PE-like materials possessing ester functionality that enables depolymerization back to the building blocks for repolymerization (*19*). However, this approach produces a limited scope of materials and yields polymers that possess high ester content and lower thermal and materials properties relative to PE. The ester functionality does not drastically impact the unit-cell identity relative to PE but does lead to decreased lamellar thickness, destabilization of the crystal lattice, and lowering of the melting transition temperature (T_m) (*20*). Although condensation polymerization of dicarboxyl aliph[illegible] oligomeric building blocks with small-mole[illegible] diols affords plastics with lower ester content for improved thermal and materials properties more comparable to PE (*21*), traditional step-growth polymerizations require precise stoichiometry matching of complimentary polymerizable groups (i.e., esters and alcohols) of the monomers to be coupled to achieve high molecular weight and technologically useful polymers. Even slight imbalances in stoichiometry result in drastically reduced polymer molecular weights and associated properties (*22*). Therefore, because of this constraint it is difficult to access polymers that possess disparate properties through a step-growth polymerization without completely changing the chemical identity of the building blocks.

We recognized that a major challenge to realizing diverse and recyclable polyolefin-like materials through a step-growth polymerization would require a polymerization methodology able to overcome the constraints of precise stoichiometry matching of chain-end groups and allow for deviations in monomer feed ratios. Notably, a step-growth polymerization of monomers or oligomers possessing identical chain-end groups would overcome this challenge and allow for modulation of the monomer feed composition to produce diverse high–molecular weight polymers with functionality designed for recyclability (Fig. 1A). The polymerization of oligomers presents the opportunity to readily produce multiblock polymers that can integrate multiple properties into a single macromolecule. For example, multiblock polyolefins with sequences of crystalline high-density PE and amorphous low-density polyolefin have both a high T_m and a low glass transition temperature (T_g), which provide strength and elasticity across a large window of operating temperatures and conditions (*23*). By contrast, the T_m of ethylene and α-olefin statistical copolymers decreases with increasing α-olefin incorporation (*23*). With this objective in mind, we were inspired by the ruthenium-catalyzed dehydrogenation polymerization of alcohols to produce polyesters (*24*). Furthermore, in the presence of hydrogen, the ruthenium complex can catalyze the depolymerization of polyesters to diol building blocks (*25*, *26*), which suggests that a system for closed-loop chemical recycling could be developed using this methodology. This proposed circular plastics life cycle would be driven by the closed-loop removal and incorporation of hydrogen, the simplest and smallest molecule possible for a condensation polymerization and a sought-after foundation for green energy (*27*). Here, we report an approach for synthesizing multiblock polyolefin-like materials constructed from hard and soft aliphatic oligomeric building blocks with highly

[1]Department of Chemistry, Colorado State University, Fort Collins, CO 80523, USA. [2]School of Advanced Materials Discovery, Colorado State University, Fort Collins, CO 80523, USA. [3]Renewable Resources and Enabling Sciences Center, National Renewable Energy Laboratory, Golden, CO 80401, USA. [4]BOTTLE Consortium, Golden, CO 80401, USA.
*Corresponding author. Email: garret.miyake@colostate.edu
†These authors contributed equally to this work.

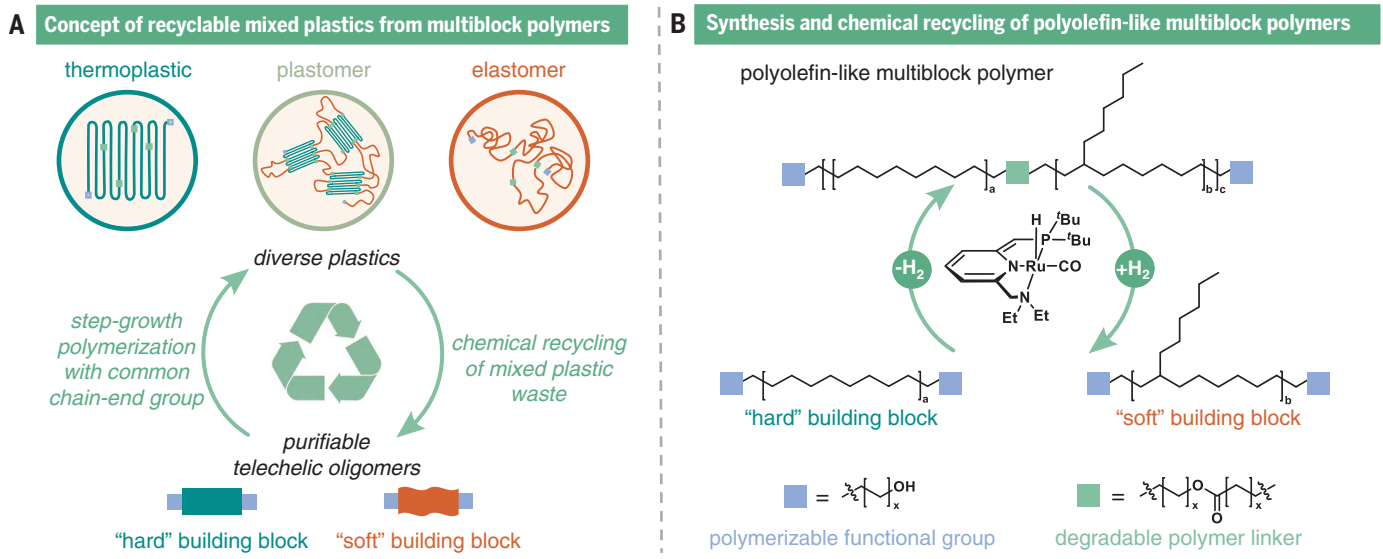


Fig. 1. Overview of chemically recyclable polyolefin-like multiblock materials with tunable properties from polymerization of hard and soft oligomers. This strategy (**A**) successfully builds on an approach for the synthesis of ester-linked polyolefin-like multiblock polymers to create a platform for closed-loop recycling of mixed plastics catalyzed by a ruthenium complex (**B**).

Table 1. Properties of multiblock polymers tuned by controlling the hard-block incorporation. See supplementary materials for details. For PEx samples, x represents the feed ratio of hard block in copolymers. Weight average molecular weight, M_w; dispersity, $Đ$; degree of crystallinity, X_c; glass transition temperature, T_g; melting transition temperature, T_m; decomposition temperature at 5% weight loss, T_{d5}; Young's modulus, E; tensile strength, σ_{UTS}; elongation at break, ε_b; tensile toughness, U_T; zero-shear viscosity, η_0. Average values for E, σ_{UTS}, ε_b, and U_T are provided with standard deviations from the mean value of four to six samples.

Sample	M_w (kDa)	$Đ$ (M_w/M_n)	Branches (per 1000C)	X_c (%)	T_g (°C)	T_m (°C)	T_{d5} (°C)	E (MPa)	σ_{UTS} (MPa)	ε_b (%)	U_T ($MJ \cdot M^{-3}$)	$\eta_0 \times 10^5$ (Pa·s)
PE0	71.7	2.4	62	0	−59.1	-	409	0.58 ± 0.05	0.041 ± 0.006	700 ± 100	0.19 ± 0.04	3.45
PE20	90.4	2.5	58	7.1	−60.0	106	406	5.4 ± 0.2	3.3 ± 0.3	1000 ± 100	27 ± 4	8.88
PE40	82.4	2.4	43	17	−50.6	109	413	15 ± 1	12.0 ± 0.7	1030 ± 30	77 ± 6	21.1
PE60	62.7	2.7	36	26	−46.5	112	421	84 ± 4	23 ± 1	710 ± 40	89 ± 8	68.7
PE80	81.0	2.5	25	43	-	117	416	279 ± 8	24 ± 2	720 ± 60	110 ± 10	21.4
PE100	80.7	2.2	0	68	-	124	415	800 ± 30	25 ± 2	800 ± 100	150 ± 30	447
HDPE	92.2	5.4	1.9	53	-	128	430	1100 ± 100	24 ± 1	530 ± 40	97± 7	2.47
LLDPE	90.4	2.8	30	32	−27.4	121	434	270 ± 10	28 ± 2	750 ± 60	130 ± 20	1.30

tunable materials properties dictated by the ratio of the two types of blocks. Notably, the different multiblock polymers can be mixed and efficiently depolymerized back to their building blocks for purification and repolymerization, demonstrating the recycling of these different plastics.

Oligomer synthesis and copolymerization

To produce the required telechelic diol-functionalized hard [**HO-HB-OH**; HB, hydrogenated poly(cyclooctene)] and soft [**HO-SB-OH**; SB, hydrogenated poly(3-hexylcyclooctene)] oligomeric blocks, we used ruthenium-mediated ring-opening metathesis polymerization (*28*) of cyclooctene or 3-hexylcyclooctene, respectively, in the presence of *cis*-hexadec-6-ene-1,16-diol as a chain-transfer agent (fig. S1). These oligomers were subsequently hydrogenated (40 bar H_2, 100°C for 36 hours), catalyzed by residual ruthenium in the unpurified product, to yield the hard **HO-HB-OH** [weight-average molecular weight (M_w) = 3.3 kDa] and soft **HO-SB-OH** (M_w = 3.0 kDa) building blocks (figs. S2 to S4). The crystallizable hard block was designed to imbue high T_m and modulus into the polymer, whereas the hexyl-functionalized soft block was designed to introduce controlled short-chain C6 branching to create noncrystallizable, elastomeric soft domains. Branching has substantial influences on polyolefin thermal and mechanical properties, primarily by reducing the allowable degree of crystallinity and increasing free volume (*29*). The oligomers were polymerized in high yields (>90%) with varying ratios of hard and soft blocks to modulate the hard content from 0 to 100% (figs. S5 to S10) by using ruthenium-catalyzed dehydrogenation polymerization to high–molecular weight polymers (M_w = 62.7 to 90.4 kDa) (figs. S11 to S14 and table S1) with high dispersity ($Đ$ > 2.2), as expected for polymers produced through a step-growth polymerization (Fig. 1B and table S2). The branching content of the polymers was determined by proton nuclear

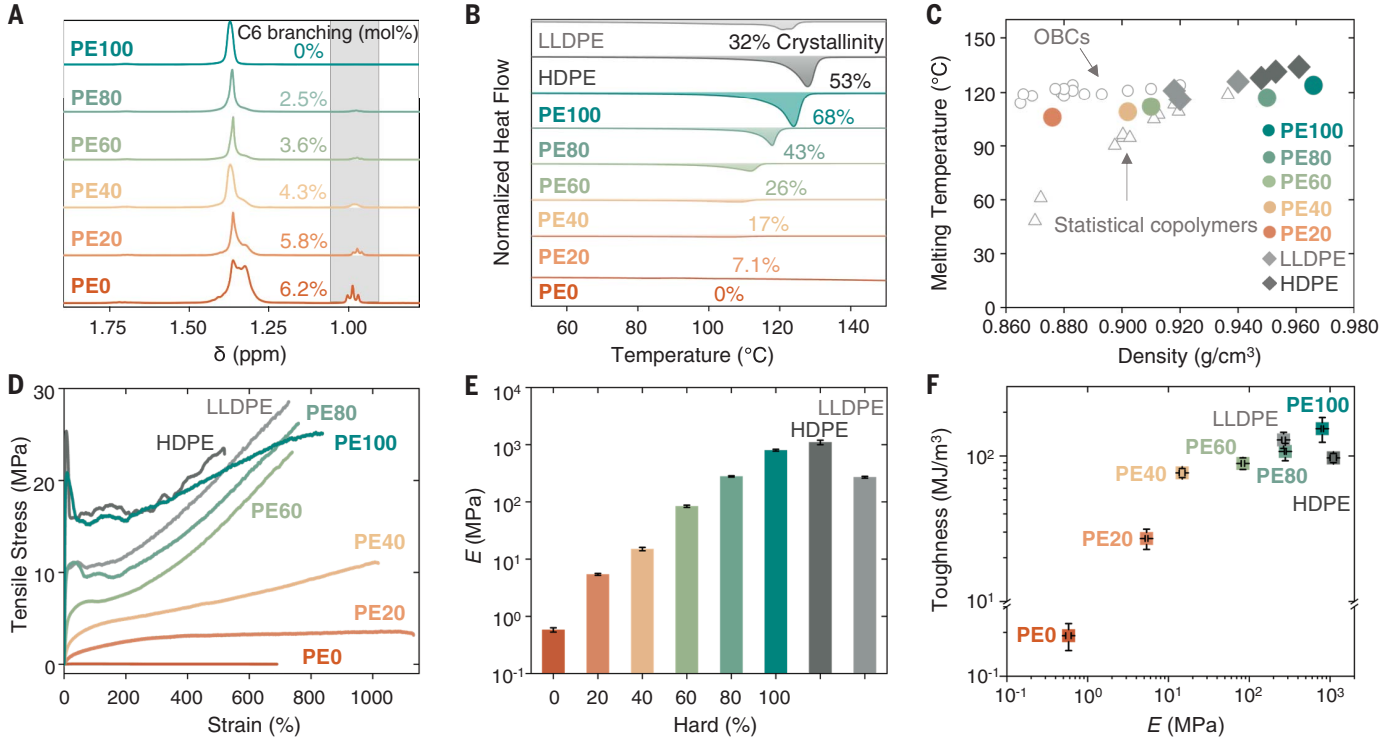


Fig. 2. Properties of multiblock polymers can be modulated over diverse regimes. (**A**) Stacked ^{1}H NMR spectra, (**B**) normalized differential scanning calorimetry traces, (**C**) melting temperature of multiblock polymers, commercial olefin block copolymers (OBCs), HDPE, LLDPE, and statistical olefin copolymers as a function of density, (**D**) representative stress-strain curves, (**E**) modulus, and (**F**) property comparison of toughness and modulus between multiblock polymers, HDPE, and LLDPE. Error bars represent standard deviations from the mean value of four to six samples.

magnetic resonance spectroscopy (^{1}H NMR) and was consistent with the block feed ratios (Fig. 2A). Although the catalyst proved difficult to remove and colored the polymers (tables S3 and S4), advantageously, depolymerization could be successfully performed with this residual catalyst (table S9, entry S22). From sustainability and practical perspectives, precious metal catalysts for commodity plastic production are undesirable, and the residual ruthenium contamination could impede use in biomedical or food packaging applications, especially those in which transparency is desired. Toward addressing this concern, lowering the catalyst concentration or using a different ruthenium complex produced considerably less colored materials (table S5 and figs. S15 and S16).

Polymer properties

The resulting polymers were thermally stable, with high decomposition temperatures at 5% weight loss (T_{d5} = 406° to 421°C; fig. S17), consistent with commercially available high-density PE (HDPE; T_{d5} = 430°C) or linear low-density PE (LLDPE; T_{d5} = 434°C) samples (Table 1). With increasing hard content, the crystallinity of the polymers increased from 0 to 68% (Fig. 2B). Polymers with hard content <80% exhibited low glass transition temperatures (T_g = −46.5° to −60.0°C; figs. S18 and S19 and table S6), whereas a high melting transition temperature (T_m = 106° to 124°C) was observed for all multiblock polymers containing hard blocks, similarly to olefin-block copolymers (OBCs). (Fig. 2C and table S7). Further examination of the crystallinity by wide-angle x-ray scattering (WAXS) showed peaks corresponding to the orthorhombic unit cell of PE (fig. S20), although polarized light optical microscopy revealed that the crystallites were smaller in size than LLDPE or HDPE samples because of the presence of the ester group (fig. S21) (*20*). Thus, the multiblock polymers possessed both a high T_m and low T_g to allow for thermoplastic and elastic properties across a wide operating temperature.

Rheology was performed to investigate the creep compliance and zero-shear viscosity (η_0) of the multiblock polymers. Generally, η_0 increased with increasing hard content from **PE0** to **PE100** (η_0 = 3.45 to 447 ×10^{5} Pa·s), indicating that the introduction of short-chain branching reduced entanglement (fig. S22 and table S8). The molecular weight variations, combined with different hard and soft compositions and statistical block sequence formations, led to imprecise correlations with these properties (*30*). In comparison with polyolefins of similar molecular weights, these multiblock polymers exhibit increased η_0, which hints at different interactions that may prove advantageous for use at temperatures approaching the T_m with reduced concern for material deformation (table S8). We postulate that this result may be due to structural effects imparted by functional group interactions from the addition of esters or from the block-polymer organization and effects due to the disparity in block sequences across varying compositions.

The mechanical properties of the multiblock polymers were investigated by uniaxial tensile testing (Fig. 2D and figs. S23 and S24). Polymer properties spanned drastic regimes, from elastomers to plastomers to thermoplastics, by increasing the hard content. Both Young's modulus (E) and tensile strength (σ_{UTS}) increased by over three orders of magnitude (E = 580 kPa for **PE0** to 800 MPa for **PE100**; σ_{UTS} = 40.5 kPa for **PE0** to 24.7 MPa for **PE100**) (Fig. 2E and fig. S25). Furthermore, all samples spanning these regimes demonstrated excellent extensibility, with the average strain at break (ε_b) for all multiblock polymers exceeding 700%, with **PE20** and **PE40** achieving >1000% strain at break. These materials also exhibited tunable tensile toughness (U_T) ranging from 0.19 to 150 MJ•m^{-3} (Fig. 2F). **PE60**, **PE80**, and **PE100** all exhibited thermoplastic behavior with yield points and strain hardening,

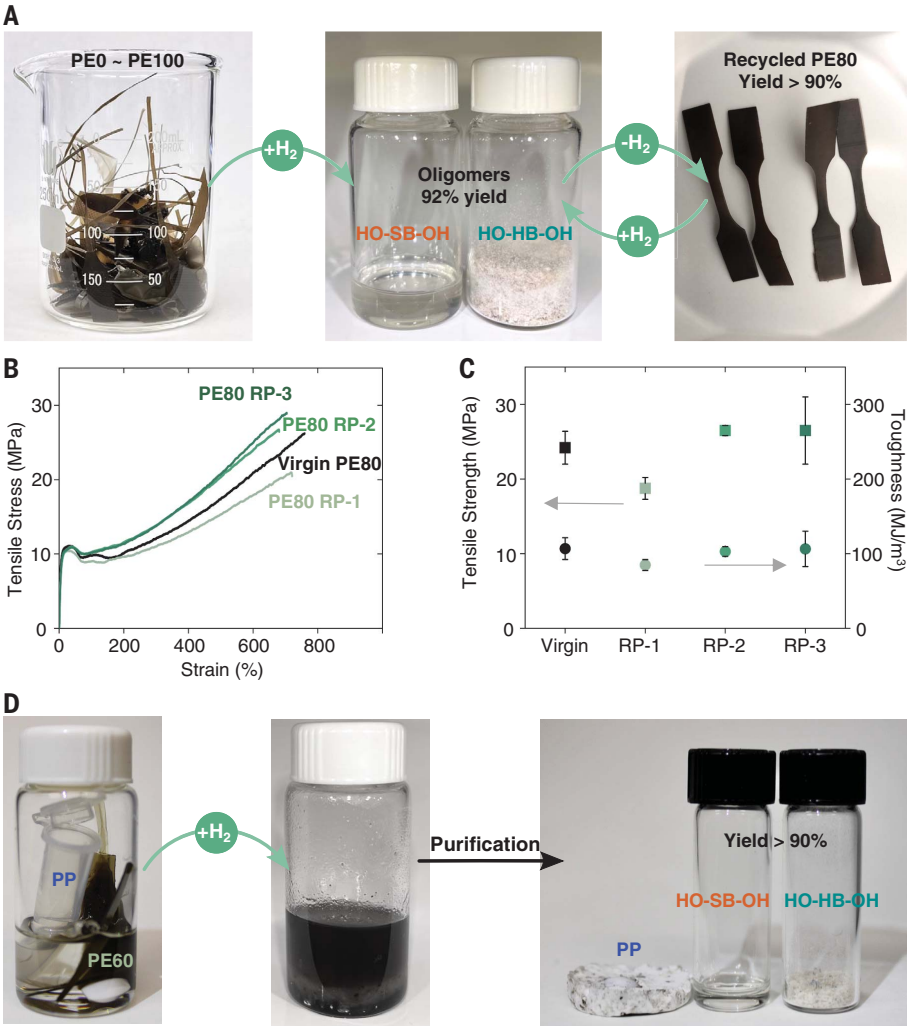


Fig. 3. Chemical recycling of multiblock polymers. (**A**) Photographs of the chemical recycling of mixed multiblock polymers. (**B**) Representative stress-strain curves of virgin **PE80** and repolymerized **PE80 RP-1**, **PE80 RP-2**, and **PE80 RP-3**. (**C**) Tensile strength and toughness of virgin **PE80** and repolymerized **PE80 RP-1**, **PE80 RP-2**, and **PE80 RP-3**. Error bars represent standard deviations from the mean value of four to five samples. (**D**) A mixture of PP and **PE60** as an example of a potential polyolefin waste stream. Photographs of mixed plastics in toluene (left), reaction after 72 hours (middle), and the separated PP, **HO-HB-PH**, and **HO-SB-OH** (right). See supplementary materials for details.

whereas a yield point was not observed for elastomeric polymers containing lower PE content. The overall thermal and materials properties of these polymers are akin to commercial OBCS. They also exhibit greater strain at break relative to commercially available HDPE and LLDPE samples tested in this work, and the toughness of **PE80** and **PE100** are comparable to or exceed that of very tough engineering plastics (>100 MJ•m^{-3}) (table S7).

Recycling studies

To complete the closed-loop chemical recycling process, the depolymerization of **PE0** was optimized to convert the polymer back to the fundamental oligomeric building blocks (fig. S26 and table S9). Although the residual ruthenium catalyst in the polymer with 40 bar H_2 at 160°C in toluene was found sufficient to catalyze the depolymerization after 24 hours (entry S22), additional catalyst was added to ensure sufficient active species in recycling studies (table S3). The closed-loop chemical recycling process of mixed-plastic waste was demonstrated; in doing so, the multiblock polymers of all compositions following tensile testing were combined for simultaneous depolymerization (>99% conversion) and the hard and soft oligomeric building blocks were separated and purified in 91.7% isolated yield with no observable signs of oligomer decomposition by ^{1}H NMR (Fig. 3A and figs. S27 and S28). The hard and soft blocks were readily separated through industrially viable selective-solvent isolation (*31*) by precipitation of the hard block and purified to remove catalyst residues to low parts-per-million levels such that ruthenium levels did not increase during subsequent polymerizations. These recovered oligomers were successfully repolymerized to **PE80**, which was subsequently depolymerized and repolymerized back into **PE80** for two additional cycles with all steps proceeding in high yields (tables S10 and S11). The molecular weights of the recycled **PE80** multiblock polymers remained consistent (M_w = 73.8 to 96.6 kDa) (figs. S29 and S30), demonstrating the robustness of this closed-loop recycling process. Notably, tensile testing of all recycled **PE80** samples revealed that the modulus, tensile strength, elongation at break, and toughness were comparable to the virgin **PE80** sample (Fig. 3, B and C, and fig. S31). We demonstrated that these multiblock polymers can be separated from technologically important isotactic polypropylene (PP) (Fig. 3C and figs. S32 and S33). Mixtures of PE and PP are difficult to separate, do not homogeneously blend, and result in materials with compromised properties, emphasizing the importance of being able to separate mixed plastics or devise strategies for compatibilization (*32*). In the presence of commercially available PP, the depolymerization of **PE60** was efficient, and the oligomers were readily separated from the PP in high yield (92%) (fig. S32).

Despite vast synthetic strategies available to produce polymers, the ability to synthesize multiblock polymers remains a challenge most commonly met through tedious sequential monomer additions (*33*). Synthesizing multiblock polymers from hard and soft aliphatic oligomers possessing identical chain-end groups creates

a platform to produce highly tunable polyolefin-like materials and their closed-loop chemical recycling process, including in the presence of other commercially important plastics. The modularity of this approach coupled with improvements focused on enhancing sustainability will further allow realization of other chemically recyclable plastics enabled by the multiblock polymer architecture for the union of otherwise unobtainable polymer compositions (*34*). Although this approach provides a promising resolution to pressing challenges of plastic recycling, we foresee that forthcoming endeavors will uncover nonprecious-metal catalyst systems that exhibit enhanced performance to enable commercial impact.

REFERENCES AND NOTES

1. R. Geyer, J. R. Jambeck, K. L. Law, *Sci. Adv.* **3**, e1700782 (2017).
2. P. Galli, G. Vecellio, *J. Polym. Sci. A Polym. Chem.* **42**, 396–415 (2004).
3. A. Chamas *et al*., *ACS Sustain. Chem. & Eng.* **8**, 3494–3511 (2020).
4. M. MacLeod, H. P. H. Arp, M. B. Tekman, A. Jahnke, *Science* **373**, 61–65 (2021).
5. B. D. Vogt, K. K. Stokes, S. K. Kumar, *ACS Appl. Polym. Mater.* **3**, 4325–4346 (2021).
6. H. Li *et al*., *Green Chem.* **24**, 8899–9002 (2022).
7. R. J. Conk *et al*., *Science* **377**, 1561–1566 (2022).
8. F. Zhang *et al*., *Science* **370**, 437–441 (2020).
9. S. Liu, P. A. Kots, B. C. Vance, A. Danielson, D. G. Vlachos, *Sci. Adv.* **7**, eabf8283 (2021).
10. X. Jia, C. Qin, T. Friedberger, Z. Guan, Z. Huang, *Sci. Adv.* **2**, e1501591 (2016).
11. A. H. Mason *et al*., *Nat. Commun.* **13**, 7187 (2022).
12. J.-B. Zhu, E. M. Watson, J. Tang, E. Y.-X. Chen, *Science* **360**, 398–403 (2018).
13. G. W. Coates, Y. D. Y. L. Getzler, *Nat. Rev. Mater.* **5**, 501–516 (2020).
14. C. Shi *et al*., *Chem* **7**, 2896–2912 (2021).
15. S. Kakadellis, G. Rosetto, *Science* **373**, 49–50 (2021).
16. F. Stempfle, P. Ortmann, S. Mecking, *Chem. Rev.* **116**, 4597–4641 (2016).
17. T. Shiono, N. Naga, K. Soga, *Makromol. Chem., Rapid. Commun.* **12**, 387–392 (1991).
18. M. Baur, F. Lin, T. O. Morgen, L. Odenwald, S. Mecking, *Science* **374**, 604–607 (2021).
19. M. Haeussler, M. Eck, D. Rothauer, S. Mecking, *Nature* **590**, 423–427 (2021).
20. M. P. F. Pepels, M. R. Hansen, H. Goossens, R. Duchateau, *Macromolecules* **46**, 7668–7677 (2013).
21. A. S. Arrington, J. R. Brown, M. S. Win, K. I. Winey, T. E. Long, *Polym. Chem.* **13**, 3116–3125 (2022).
22. J. K. Stille, *J. Chem. Educ.* **58**, 862–866 (1981).
23. D. J. Arriola, E. M. Carnahan, P. D. Hustad, R. L. Kuhlman, T. T. Wenzel, *Science* **312**, 714–719 (2006).
24. D. M. Hunsicker, B. C. Dauphinais, S. P. Mc Ilrath, N. J. Robertson, *Macromol. Rapid Commun.* **33**, 232–236 (2012).
25. E. M. Krall *et al*., *Chem. Commun.* **50**, 4884–4887 (2014).
26. S. Westhues, J. Idel, J. Klankermayer, *Sci. Adv.* **4**, eaat9669 (2018).
27. G. W. Crabtree, M. S. Dresselhaus, M. V. Buchanan, *Phys. Today* **57**, 39–44 (2004).
28. S. Hilf, A. F. M. Kilbinger, *Nat. Chem.* **1**, 537–546 (2009).
29. M. Gahleitner, *Prog. Polym. Sci.* **26**, 895–944 (2001).
30. F. J. Stadler, V. Karimkhani, *Macromolecules* **44**, 5401–5413 (2011).
31. T. W. Walker *et al*., *Sci. Adv.* **6**, eaba7599 (2020).
32. J. M. Eagan *et al*., *Science* **355**, 814–816 (2017).
33. F. S. Bates *et al*., *Science* **336**, 434–440 (2012).
34. C. J. Hawker, K. L. Wooley, *Science* **309**, 1200–1205 (2005).

ACKNOWLEDGMENTS

The authors thank the Analytical Resources Core (RRID: SCR_021758) at Colorado State University for instrument access and training; C. Josephitis and M. Nicki for their assistance in early hydrogenation experiments; B. Newell and M. Mailhot for their generous assistance with instrument configuration and training; and E. Quinn, M. Price, and X. Liu for insightful discussions throughout the experimental design. **Funding:** The research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award no. R35GM144356. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We also acknowledge support from RePLACE (Redesigning Polymers to Leverage a Circular Economy) funded by the Office of Science of the US Department of Energy through award no. DE-SC0022290. Funding for N.A.R. and J.M. was provided in part by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, Advanced Manufacturing Office (AMO), and Bioenergy Technologies Office (BETO). This work was performed as part of the BOTTLE™ Consortium and was supported by AMO and BETO under contract no. DE-AC36-08GO28308 with the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC. **Author contributions:** Conceptualization: G.M.M., Y.Z., K.L.H., and E.M.R.; Methodology: G.M.M., Y.Z., K.L.H., and E.M.R.; Investigation: Y.Z., E.M.R., K.L.H., Z.H., J.M., and N.A.R.; Visualization: Y.Z., E.M.R., K.L.H., and G.M.M.; Funding acquisition: G.M.M. and N.A.R.; Project administration: G.M.M., Y.Z., and E.M.R.; Supervision: G.M.M.; Writing – original draft: G.M.M.; Writing – review and editing: Y.Z., E.M.R., K.L.H., and G.M.M. G.M.M. and Y.Z. conceived and designed the research. Y.Z., K.L.H., E.M.R., and Z.H. performed synthesis. Y.Z., E.M.R., J.M., and N.A.R. performed characterization. Y.Z. and E.M.R. analyzed data. G.M.M. wrote the manuscript and Y.Z., E.M.R., and K.L.H. wrote the supplemental materials. **Competing interests:** The Colorado State University Research Foundation with Y.Z., E.M.R., K.L.H., and G.M.M. have filed a provisional patent application on this technology. **Data and materials availability:** All data are available in the main text or the supplementary materials. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adh3353
Materials and Methods
Supplementary Text
Figs. S1 to S33
Tables S1 to S11
References (*35*–*52*)

Submitted 24 February 2023; resubmitted 19 June 2023
Accepted 6 September 2023
10.1126/science.adh3353

EPISTASIS

Idiosyncratic and dose-dependent epistasis drives variation in tomato fruit size

Lyndsey Aguirre[1], Anat Hendelman[2], Samuel F. Hutton[3], David M. McCandlish[2]*, Zachary B. Lippman[1,2,4]*

Epistasis between genes is traditionally studied with mutations that eliminate protein activity, but most natural genetic variation is in cis-regulatory DNA and influences gene expression and function quantitatively. In this study, we used natural and engineered cis-regulatory alleles in a plant stem-cell circuit to systematically evaluate epistatic relationships controlling tomato fruit size. Combining a promoter allelic series with two other loci, we collected over 30,000 phenotypic data points from 46 genotypes to quantify how allele strength transforms epistasis. We revealed a saturating dose-dependent relationship but also allele-specific idiosyncratic interactions, including between alleles driving a step change in fruit size during domestication. Our approach and findings expose an underexplored dimension of epistasis, in which cis-regulatory allelic diversity within gene regulatory networks elicits nonlinear, unpredictable interactions that shape phenotypes.

Epistasis analysis is an essential tool for discovering functional relationships between genes. At its simplest, an epistatic interaction is determined by testing whether the phenotypic effect from one gene mutation modifies (e.g., suppresses or enhances) the phenotypic effect of another (*1*, *2*). Historically, epistasis studies have relied on mutations with strong effects on protein function and phenotype, typically obtained from natural mutants or laboratory mutagenesis experiments (*1*–*4*). Recently, high-throughput engineering and the combination of gene deletions in yeast have allowed for the characterization of global interaction networks (*5*–*10*). Although these and related studies, including those now leveraging genome-editing technologies in more complex systems (*11*–*15*), can dissect epistasis at scale, they do not address how cis-regulatory mutations—which are pervasive in genomes and responsible for the majority of functional variation in organisms (*16*–*19*)—affect epistatic relationships and the phenotypes they control.

Compared with protein-coding mutations, cis-regulatory mutations more often produce graduated effects on gene function that alter expression level or timing (*16*, *20*, *21*). Across species, natural variation in gene expression is predominantly associated with regulatory sequences of the differentially expressed genes (*16*, *19*, *22*), and cis-regulatory variants are the primary contributors to phenotypic diversity (*16*, *18*). However, despite the critical functional role of these variants, few studies have explored epistatic relationships in the context of cis-regulatory variation (*5*, *10*, *23*), and none have done so in depth. Because of limited allelic variation at known interacting genes and inadequate quantitative phenotyping power in most model systems, we lack an understanding of how this widespread genetic variation affects the form and magnitude of epistasis.

We addressed this knowledge gap by taking advantage of the *CLAVATA-WUSCHEL* (*CLV-WUS*) gene regulatory circuit in plants (*24*). *CLV-WUS* controls stem-cell proliferation in small groups of cells at shoot apices called meristems, which enable the continuous development of new tissues and organs during postembryonic growth (*24*). Using tomato as a model, we asked how previously documented epistatic interactions in this circuit are affected by replacing one critical gene, *CLAVATA3* (*CLV3*), with a wide range of stronger and weaker cis-regulatory alleles.

CLV3 encodes a small signaling peptide that restricts stem-cell proliferation and meristem size by repressing *WUS*, a stem-cell–promoting homeobox transcription factor gene (*24*). In a negative feedback loop, *WUS* suppresses its own expression by activating *CLV3* to restrict stem-cell proliferation and maintain meristem size throughout development (*24*). Epistasis between *CLV3* and *WUS* was first established by using mutants in the model *Arabidopsis thaliana* (*25*), and our previous CRISPR-Cas9 mutagenesis of the tomato orthologs has shown that this relationship is conserved (*26*–*28*). In both systems, meristem growth in *wus* mutants ceases during vegetative development, resulting in a failure to develop flowers and fruits. Conversely, meristems of *clv3* mutants become greatly enlarged, leading to more flowers, fruits, and their associated organs, including seed compartments known as locules. In a classical suppression epistatic relationship, *wus* mutations completely mask *clv3* phenotypes (*clv3 wus* double mutants are indistinguishable from *wus* single mutants). Tomato also features an additional layer of epistasis involving a paralog of *SlCLV3* (*Solanum lycopersicum*, denoted by *Sl*) in the *CLV3/EMBRYO-SURROUND* *REGION* (*CLE*) gene family, *SlCLE9* (*27*). *SlC* is an ancient paralog, whose natural allelic state in wild and domesticated tomatoes is a partial loss of function (hypomorphic) resulting from changes in both its protein sequence and cis-regulatory control (*27*, *29*). Whereas null mutants of *Slcle9* are indistinguishable from wild-type (WT) plants, *Slclv3* is strongly enhanced by *Slcle9*, demonstrating a canonical unequal redundancy (*30*) epistatic relationship between these paralogs.

Although conventional protein-coding null mutations were used to characterize these epistatic relationships, two natural cis-regulatory alleles of *SlWUS* and *SlCLV3* are also known to exhibit a strong epistatic interaction (*26*). In fact, this interaction played an important role in the expansion of fruit size through an increase in locule number that occurred during tomato domestication (*26*, *31*). Specifically, the ancestral state of tomato, which is maintained in many cultivated genotypes, is to produce fruits with two or three locules (Fig. 1A). A quantitative trait locus (QTL) allele known as *locule number* (*lc*) then emerged in the progenitor of modern tomatoes (*31*). This allele disrupts a repressor element downstream of *SlWUS* ($Slwus^{lc}$), leading to a weak gain of function and a slight increase of ~10% in the number of three-locule fruits (*26*). Subsequently, another QTL allele, *fasciated* (*fas*), arose in the form of an inversion that reduces the activity of the *SlCLV3* promoter ($Slclv3^{fas}$) (*26*, *31*), resulting in twice as many locules as are found in WT plants ($SlCLV3^{FAS}$). The combination of these cis-regulatory alleles in homozygous double-mutant plants ($Slclv3^{fas}$ $Slwus^{lc}$) produces an enhanced (synergistic) epistatic effect on locule number that surpasses their combined individual effects (Fig. 1A) (*26*). Thus, the emergence of $Slclv3^{fas}$ in the context of the preexisting $Slwus^{lc}$ background is thought to have been a key step in the increase in fruit size observed during tomato domestication (*31*). However, additional cis-regulatory alleles of the *SlCLV3* locus exist (*32*, *33*), and it remains an open question whether this synergistic interaction is specific to $Slclv3^{fas}$ or whether other cis-regulatory alleles of this gene with varying allelic strengths would also exhibit epistatic enhancement with $Slwus^{lc}$.

Epistasis across an allelic series of cis-regulatory mutations

Using natural alleles to investigate the impact of cis-regulatory allelic diversity on epistatic interactions in any system is challenging because of their varied genetic backgrounds and our limited understanding of their phenotypic effects. Previously, we used CRISPR-Cas9 to engineer cis-regulatory deletion mutations that overlapped with the disrupted cis-regulatory sequences of $Slwus^{lc}$ and $Slclv3^{fas}$, resulting

[1]Cold Spring Harbor Laboratory, School of Biological Sciences, Cold Spring Harbor, NY, USA. [2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [3]Gulf Coast Research and Education Center, University of Florida, Wimauma, FL, USA. [4]Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA.
*Corresponding author. Email: mccandlish@cshl.edu (D.M.M.); lippman@cshl.edu (Z.B.L.)

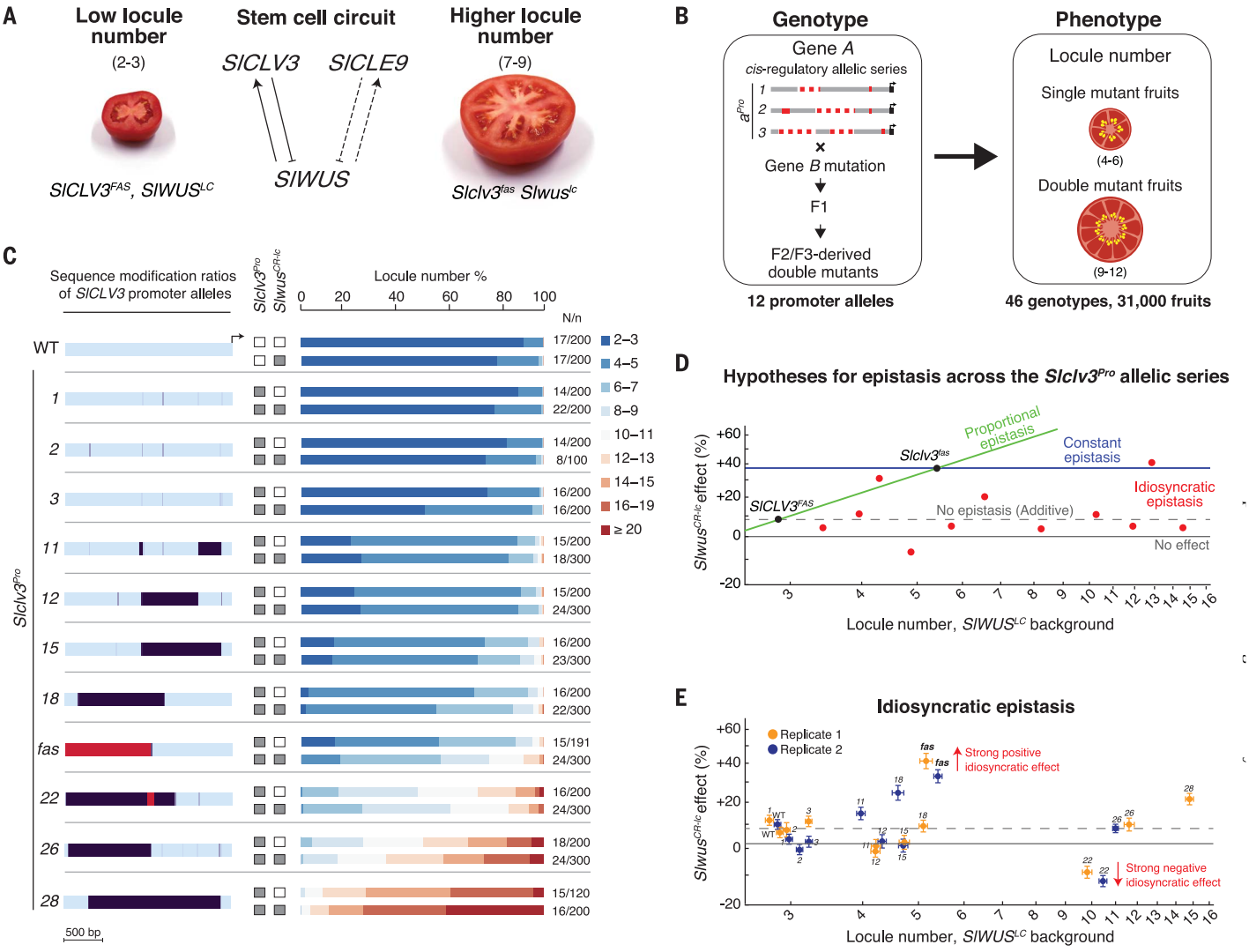


Fig. 1. A promoter allelic series of the fruit-size gene *SlCLV3* reveals idiosyncratic epistasis. (**A**) The *SlWUS-SlCLV3* circuit and the paralog *SlCLE9* control locule number. Images show fruits of WT plants (left) and of *Slclv3*^{fas} *Slwus*^{lc} double mutants (right). Dashed lines and numbers indicate locules. (**B**) Experimental design. (**C**) Heatmap of *SlCLV3* promoter region encompassing 11 *Slclv3* promoter (*Slclv3*^{Pro}) alleles. Purple intensity in 20 base-pair (bp) windows indicates ratios of sequence change relative to WT (light blue). Red intensity indicates inversion. Stacked bar charts are percentage of fruits having each locule-number range. White and gray boxes indicate WT and mutant genotype for each gene, respectively. N/n, total replicated plants (N) and total replicated fruits (n). (**D**) Epistasis models between *Slwus*$^{CR\text{-}lc}$ and the *Slclv3*^{Pro} alleles, depicted by plotting percent change of double mutants against mean log locule numbers of *Slclv3*^{Pro} mutants. The combined effect of *Slwus*^{lc} and *Slclv3*^{fas} is indicated. (**E**) *Slwus*$^{CR\text{-}lc}$ effect on mean log locule number (*Slwus*$^{CR\text{-}lc}$ *Slclv3*^{Pro} genotypes compared with *SlWUS*^{LC} *Slclv3*^{Pro} genotypes), plotted against mean log locule number of the corresponding *SlWUS*^{LC} *Slclv3*^{Pro} genetic background (error bars indicate ±1 SE). Data are from two replicated trials, except for *Slclv3*$^{Pro\text{-}28}$ (see also fig. S2A and tables S2 and S3). Red arrows show strongest idiosyncratic effects, including positive synergism between *Slclv3*^{fas} and *Slwus*^{lc}.

in mimics of their individual effects in the same genetic background (*26*, *28*). In the same experiment, we engineered an additional 28 *Slclv3* promoter alleles (*Slclv3*^{Pro}), resulting in a continuum of locule-number variation ranging from subtle increases in the proportion of three-locule fruit to strong *Slclv3* null-like effects shown in fruits that on average contain more than 15 locules (*28*). Leveraging this genetic resource and its power to quantify locule number over a wide phenotypic range, we tested whether the *Slwus*^{lc} mimic (*Slwus*$^{CR\text{-}lc}$) consistently enhances the effects of *Slclv3*^{Pro} cis-regulatory alleles to the same degree as with *Slclv3*^{fas}, or whether epistatic interactions are dependent on the allelic strength and/or specific identity of the *Slclv3*^{Pro} alleles.

From the pool of available *Slclv3*^{Pro} alleles, we selected 12 that represent the full spectrum of locule-number variation, including *Slclv3*^{fas}, and demonstrated that their homozygous mutant effects are reproducible across multiple years and environments (fig. S1A and table S3). This resource allowed us to measure how the magnitude of the epistatic interaction with *Slwus*$^{CR\text{-}lc}$ changes across this allelic series of cis-regulatory mutants (Fig. 1C). To evaluate the combined effects of *Slclv3*^{Pro} and *Slwus*$^{CR\text{-}lc}$ cis-regulatory alleles, we created all possible double-mutant combinations in the same genetic background as that of the single mutants (Fig. 1B and fig. S1, A and B). We then quantified locule numbers from all 2 × 12 = 24 genotypes, including WT and single mutants, across two replicated experiments (Fig. 1, B and C, and fig. S2A).

We considered several specific hypotheses on how the magnitude of this epistatic interaction (table S2) might change as a function of

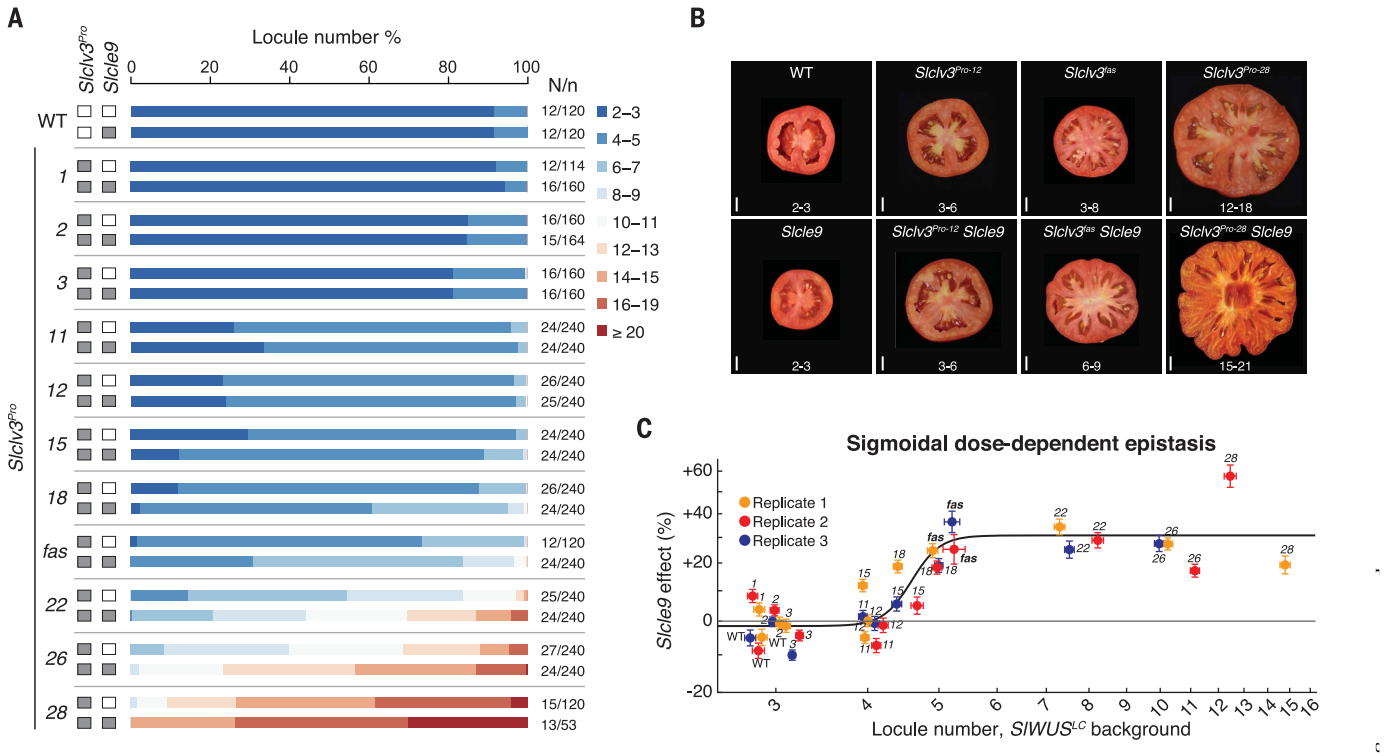


Fig. 2. The compensating paralog *SlCLE9* interacts with *SlCLV3* in a sigmoidal dose-dependent epistasis relationship. (**A**) Stacked bar charts show percentage of total fruits for each locule number range of *Slclv3*^{Pro} single- and *Slclv3*^{Pro} *Slcle9* double-mutant alleles. White and gray boxes indicate WT and mutant genotype for each gene, respectively. N/n, total replicated plants (N) and total replicated fruits (n). (**B**) Representative fruit images and locule-number quantification (mean ±1 SD) showing the effect of *Slcle9* on locule number in WT plants and the *Slclv3*^{Pro} mutants. Scale bars, 1 cm. (**C**) *Slcle9* effect on mean log locule number (*Slcle9 Slclv3*^{Pro} double mutants as compared with *SlCLE9 Slclv3*^{Pro} single mutants), plotted against the mean log locule number of the corresponding *SlCLE9 Slclv3*^{Pro} genetic background (error bars indicate ±1 SE). Black line indicates the maximum-likelihood fit for the sigmoid model. Data are from three replicate trials (see also fig. S2B and tables S2 and S3).

cis-regulatory allelic strength: the absence of epistasis from *Slwus*$^{CR\text{-}lc}$ (i.e., additivity), as well as three modes of epistasis across the *Slclv3*^{Pro} allelic series: proportional, constant, and idiosyncratic (Fig. 1D). In proportional epistasis (also known as the multilinear model) (*34*), the *Slwus*$^{CR\text{-}lc}$ effect scales linearly with *Slclv3*^{Pro} allelic strength, whereas in constant epistasis, the *Slwus*$^{CR\text{-}lc}$ effect is the same for each mutant allele. By contrast, idiosyncratic epistasis is allele specific in that the *Slwus*$^{CR\text{-}lc}$ effect varies, potentially in either positive or negative directions, depending on the *Slclv3*^{Pro} mutant background (*35*, *36*).

To test these hypotheses, we built a nested family of models and fit them to the log-transformed data using maximum likelihood (supplementary materials). This analysis found that although neither the constant nor proportional epistasis models provided a better fit than the additive model (likelihood-ratio test, $P = 0.88$ and $P = 0.32$, respectively), the additive, constant epistasis, and proportional epistasis models could all be rejected in favor of the idiosyncratic epistasis model (likelihood-ratio test, $P < 0.0001$ against all simpler models). Thus, the effect of *Slwus*$^{CR\text{-}lc}$ across the *Slclv3*^{Pro} allelic series is neither constant nor a simple function of allelic strength but rather varies substantially in an allele-specific manner (Fig. 1E). A notable example is *Slclv3*$^{Pro\text{-}22}$. Whereas this single mutant displays higher locule numbers than both the *Slclv3*^{fas} and *Slclv3*^{fas} *Slwus*$^{CR\text{-}lc}$ genotypes, counter to expectations, in the background of *Slclv3*$^{Pro\text{-}22}$, *Slwus*$^{CR\text{-}lc}$ actually decreases locule number, constituting a strong negative idiosyncratic effect (Fig. 1, C to E). Moreover, our analysis also shows that the strong positive idiosyncratic effect from *Slwus*^{lc} on the *Slclv3*^{fas} background was not observed with any other *Slclv3*^{Pro} alleles (Fig. 1E). Thus, the combined effect on locule number from *Slclv3*^{fas} and *Slwus*^{lc} played a distinct and critical role in enhancing fruit size during domestication, beyond what their individual effects could achieve.

The idiosyncratic epistasis between *Slwus*$^{CR\text{-}lc}$ and a subset of specific *Slclv3*^{Pro} alleles was surprising given the continuous phenotypic variation produced across the *Slclv3*^{Pro} allelic series. This raised the question of whether such unpredictability would be recapitulated with mutations of *SlCLE9,* which enhance the effects of both the *Slclv3* null mutation and the *Slclv3*^{fas} cis-regulatory mutation (*27*). Notably, similar to what occurs in *Slclv3* null alleles, the expression of *SlCLE9* is up-regulated in *Slclv3*^{fas} mutant meristems, although to a lesser degree (*27*). We confirmed this result and showed further that, overall, across the *Slclv3*^{Pro} allelic series, *SlCLE9* expression increases when *SlCLV3* expression decreases (fig. S3) as one moves from low to high locule-number alleles. These observations suggested that, unlike *Slwus*$^{CR\text{-}lc}$, which produced idiosyncratic epistasis when combined with the *Slclv3*^{Pro} allelic series, *Slcle9* could progressively enhance locule numbers across the allelic series, which would support proportional epistasis (Fig. 1D).

Using the same *Slclv3*^{Pro} mutants and approach as we did for *Slwus*$^{CR\text{-}lc}$ (Fig. 2A and fig. S1C), we unexpectedly found that for *Slcle9*, all of the simpler models were again rejected in favor of the idiosyncratic epistasis model (likelihood-ratio test, $P < 0.0001$ against all simpler models). However, unlike the findings for *Slwus*$^{CR\text{-}lc}$, in which the allele-specific effects varied substantially between phenotypically similar genetic backgrounds, in *Slcle9*, the additive model could be rejected in favor of both the constant and proportional epistasis

models (likelihood-ratio test, $P < 0.0001$ for both models), and examination of the estimated epistatic effects between all single- and double-mutant pairs (table S2) suggested that the *Slcle9* effect varied in a threshold-like manner as a function of $Slclv3^{Pro}$ allelic strength. In particular, whereas *Slcle9* had only a minimal effect on locule in the weaker $Slclv3^{Pro}$ backgrounds (which express *SlCLV3* at near-WT levels) (fig. S3), a larger effect emerged in the stronger, higher locule backgrounds where *SlCLV3* is expressed at a substantially lower level (fig. S3), including $Slclv3^{fas}$ and the near null mutant $Slclv3^{28}$ (Fig. 2B). On the basis of these observations, we fit an additional model in which the *Slcle9* effect increases as a sigmoid function of the strength of the $Slclv3^{Pro}$ background (Fig. 2C). Although the idiosyncratic epistasis model still provided a better fit to the data (likelihood-ratio test, $P < 0.0001$), the sigmoid model provided a better fit than either the constant or proportional epistasis models (likelihood-ratio test, $P < 0.0001$ against both simpler models). Moreover, if we consider the epistatic variance in log locule number as the fraction of the variance that is accounted for by the idiosyncratic epistasis model but not by the additive model, we find that the sigmoid model captures the vast majority of this variance (90.0%, table S2). We thus conclude that although there is a statistically significant idiosyncratic component to the *Slcle9* effect, the overall pattern is a dose-dependent saturating relationship, in which the effect of *Slcle9* is negligible until a critical $Slclv3^{Pro}$ allelic strength (critical degree of *SlCLV3* disruption) is reached. Above this threshold, the effect of *Slcle9* increases and eventually reaches an approximately constant level of enhancement in stronger $Slclv3^{Pro}$ backgrounds.

Higher-order mutant combinations reveal additional idiosyncrasy

Although our findings show that the effects of *Slcle9* null mutants have a sigmoid epistasis relationship across the $Slclv3^{Pro}$ allelic series, modern genotypes typically also carry $Slwus^{lc}$ (*31*). To evaluate whether this pattern is maintained in the presence of $Slwus^{CR\text{-}lc}$, we constructed and phenotyped a combinatorially complete set of triple mutants using a subset of five mutant $Slclv3^{Pro}$ alleles with a wide range of allelic strengths (Fig. 3A, $6 \times 2 \times 2 = 24$ total genotypes). We found previously unknown and unpredicted epistatic interactions in these higher-order mutants that were not present in the double mutants. Although the effect of *Slcle9* on locule number is negligible in WT, $Slwus^{CR\text{-}lc}$, and weak $Slclv3^{Pro}$ mutant backgrounds, locule number was enhanced by *Slcle9* in all triple mutants, including with the weak $Slclv3^{Pro\text{-}2}$ allele (Fig. 3A and fig. S2C), and the *Slcle9* effect broadly increased and approached saturation at approximately the level predicted by the sigmoid model (Fig. 3B). Although our previous analyses showed that $Slwus^{CR\text{-}lc}$ had a strong positive and negative idiosyncratic influence on the effects of $Slclv3^{fas}$ and $Slclv3^{Pro\text{-}22}$, respectively (Fig. 1E), we did not observe strong idiosyncratic epistasis with *Slcle9* and these alleles, even though $Slwus^{CR\text{-}lc}$ was present in the backgrounds of the triple mutants (Fig. 3 and table S2). By contrast, we observed a substantial reversal of the *Slcle9* effect on $Slclv3^{Pro\text{-}11}$ in the presence of $Slwus^{CR\text{-}lc}$, in which locule number is actually decreased, instead of increased, by the *Slcle9* mutation (Fig. 3B). Consistent with this idiosyncratic effect, the constant and proportional epistasis models were rejected in favor of the idiosyncratic epistasis model (likelihood-ratio test, $P < 0.0001$ against both simpler models). Taken together, these findings demonstrate that the predictability of epistatic effects and phenotypic outcomes in two-way interactions can be altered in higher-order allelic combinations.

Discussion

Cryptic background mutations, which have subtle or no effect on phenotype (*37*), are pervasive in genomes, and despite little knowledge about their underlying genes, alleles, and mechanisms, these mutations are widely recognized as critical factors that shape the evolutionary trajectories of traits under both natural and artificial selection (*2*, *38*–*40*). Our observations expose the dynamic role played by epistasis among the natural and cryptic alleles of these genes during tomato domestication. The natural hypomorphic *SlCLE9* allele preexisted as a cryptic variant in the genome of the wild progenitor of tomato (*27*, *29*) and was followed by $Slwus^{lc}$, whose subtle influence on locule number likely also persisted cryptically (*31*). Consequently, the later emergence of $Slclv3^{fas}$ would have immediately triggered a positive idiosyncratic epistatic interaction with $Slwus^{lc}$, wherein these $Slclv3^{fas}$ mutants displayed a marked increase in locule number that they would not have shown in the absence of these preceding mutations. Thus, the fortuitous *SLCLV3* cis-regulatory allele responsible for the initial and most consequential step in enhancing fruit size by increasing locule number during domestication appears to have had its quantitative effect arise from the combination of an unpredictable idiosyncratic interaction with the cryptic gain-of-function $Slwus^{lc}$ allele and from alleviation of dose-dependent suppression by the cryptic hypomorphic *SlCLE9*.

The idiosyncratic epistatic effects that we observe here are presumably driven by allele-specific differences in the composition and location of regulatory elements within the *SlCLV3* promoter. However, identifying the causative regulatory elements is difficult because each mutant allele typically disrupts dozens of transcription-factor binding sites (*28*) and because the regulatory architecture of meristem development remains incompletely understood (*24*). In light of the remarkable complexity of epistatic interactions originating from a limited number of background mutations and a one-dimensional array of allelic strength, our findings hold ramifications for other organisms and phenotypes in both natural genetic contexts and genetic engineering. Gene regulatory networks are the foundation of biological systems (*41*, *42*), and these networks depend on intricate signaling and feedback mechanisms—encompassing both positive and negative regulation—between genes and their protein products, often involving paralogs engaged in asymmetrical redundancy relationships (*3*, *30*, *43*). Notably, the redundancy relationship between *SlCLV3* and *SlCLE9* is based on a widespread transcriptional compensation mechanism (*27*, *29*, *43*, *44*), suggesting that similar saturating dose-dependent epistatic interactions are likely to be ubiquitous. However, varying allelic states of redundant paralogs could affect the form of dose-dependent relationships. For example, *SlCLE9* orthologs differ across Solanaceae crops, from the more potent partner of the *SlCLV3* ortholog in groundcherry to the complete loss of this gene in eggplant (*29*). These varying allelic states are important to consider when designing editing strategies to increase locule number. Likewise, how epistasis is transformed across an allelic series could also be influenced by environmental conditions. We found that the phenotypic effects of both coding and regulatory *SlCLV3* mutations are typically not substantially affected by the environment (*28*), and although the patterns of epistasis observed in our study might have some dependence on environment, the genotype-specific locule-number distributions remained remarkably consistent across different field seasons and locations (fig. S2). Applying methods similar to those used here provides a path to determine the form of these interactions for other organisms, traits, and environments, which would facilitate the fine-tuning of phenotypes in a controlled and quantitative manner.

It is important to acknowledge, however, that the predictability of outcomes when engineering new alleles and allelic combinations may be influenced by idiosyncratic interactions with other background mutations (*2*, *45*, *46*). Indeed, our observation of a previously unidentified idiosyncratic effect in the $Slclv3^{Pro}$ $Slwus^{lc}$ *Slcle9* triple mutants, which was not present in the $Slclv3^{Pro}$ $Slwus^{lc}$ double mutants, underscores how predictability of effects from engineered alleles may decay in increasingly divergent genetic backgrounds. A related issue is that natural alleles responsible for phenotypic differences between genotypes and species, which are being increasingly revealed through pangenomics (*32*, *33*, *47*, *48*),

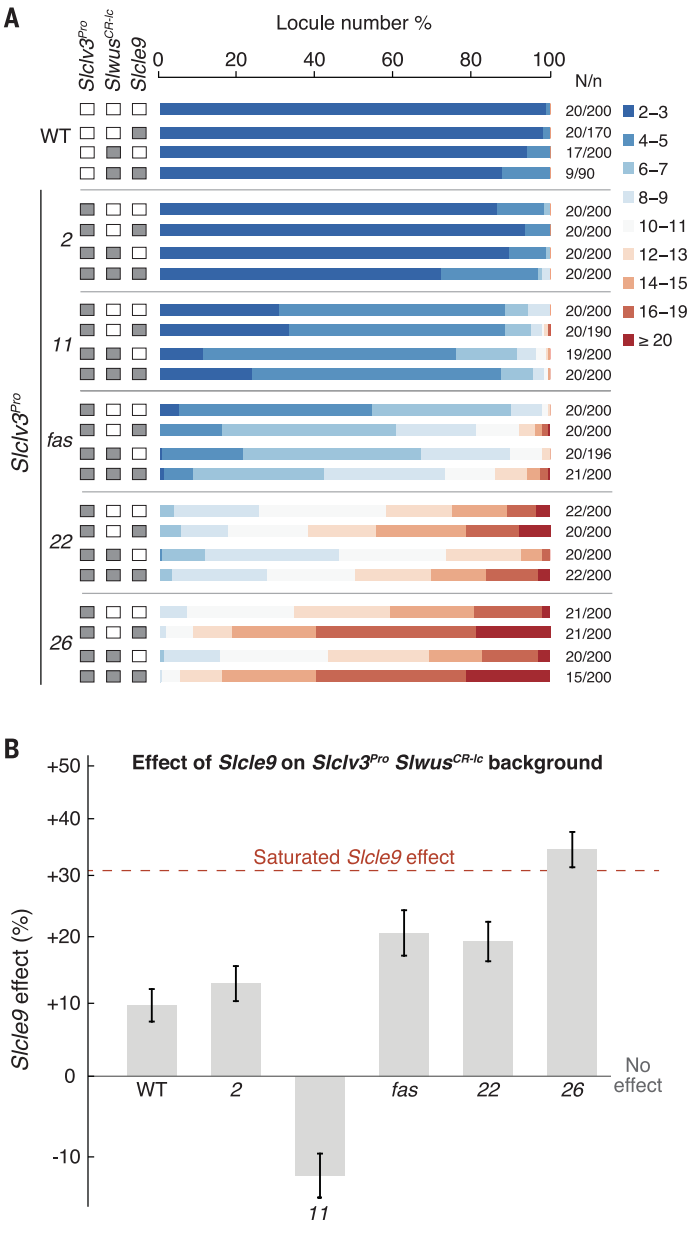


Fig. 3. Loss of *SlCLE9* imposes unpredicted idiosyncratic effects on *Slclv3*^{Pro} *Slwus*$^{CR\text{-}lc}$ backgrounds. (**A**) Stacked bar charts show percentage of total fruits for each locule-number range of WT genotype and all indicated single-, double-, and triple-mutant genotypes. White and gray boxes indicate WT and mutant genotype for each gene, respectively. N/n, total replicated plants (N) and total replicated fruits (n). (**B**) *Slcle9* effect on the log mean locule number (*Slcle9 Slwus*$^{CR\text{-}lc}$ *Slclv3*^{Pro} triple mutants as compared with the *SlCLE9 Slwus*$^{CR\text{-}lc}$ *Slclv3*^{Pro} double mutants) in the indicated *SlCLE9 Slwus*$^{CR\text{-}lc}$ *Slclv3*^{Pro} double-mutant background (error bars indicate ±1 SE). Notice the strong negative idiosyncratic epistasis in the *Slclv3*$^{Pro\text{-}11}$ *Slwus*$^{CR\text{-}lc}$ background. The black line indicates no effect, and the red dashed line indicates the saturated effect of *Slcle9* on *Slclv3*^{Pro}, on the basis of our previously fit sigmoid model (see also Fig. 2C, fig. S2C, and tables S2 and S3).

may be enriched for idiosyncratic effects resulting from the action of natural or artificial selection (*49*, *50*), as seen with *Slclv3*^{fas} and *Slwus*^{lc}. More broadly, the expected degree of variability in epistatic interactions displayed by different alleles at the same locus, how these epistatic interactions are transformed as a function of allelic strength, and whether these patterns differ between natural versus artificial alleles and regulatory versus coding sequences remain as open questions. Although we have shown that our *Slclv3*^{Pro} allelic series interacts differently with *Slwus*^{lc} (idiosyncratically) versus *Slcle9* (a systematic, dose-dependent response), it will be informative to investigate whether other allelic series will exhibit consistent or distinct patterns of epistatic interaction when the same allelic series is paired with different epistatic partners. Systematic mapping of predictable epistatic interactions, striving to minimize and perhaps leverage potential idiosyncratic effects, represents a key challenge in present and future endeavors to modify, correct, and optimize traits in agriculture and human health.

REFERENCES AND NOTES

1. T. F. C. Mackay, *Nat. Rev. Genet.* **15**, 22–33 (2014).
2. T. B. Sackton, D. L. Hartl, *Cell* **166**, 279–287 (2016).
3. B. Lehner, *Trends Genet.* **27**, 323–331 (2011).
4. R. F. Campbell, P. T. McGrath, A. B. Paaby, *Trends Genet.* **34**, 883–898 (2018).
5. A. N. Nguyen Ba *et al.*, *eLife* **11**, e73983 (2022).
6. N. Bosch-Guiteras, J. van Leeuwen, *Curr. Opin. Genet. Dev.* **76**, 101963 (2022).
7. M. Costanzo *et al.*, *Science* **353**, aaf1420 (2016).
8. M. S. Johnson, M. M. Desai, *eLife* **11**, e76491 (2022).
9. E. Caudal *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2204206119 (2022).
10. R. M. L. Ang, S.-A. A. Chen, A. F. Kern, Y. Xie, H. B. Fraser, *Cell Genomics* **3**, 100260 (2023).
11. T. M. Norman *et al.*, *Science* **365**, 786–793 (2019).
12. M. A. Horlbeck *et al.*, *Cell* **174**, 953–967.e22 (2018).
13. J. Shi *et al.*, *Nat. Biotechnol.* **33**, 661–667 (2015).
14. H. J. Liu *et al.*, *Plant Cell* **32**, 1397–1413 (2020).
15. Y. Hu *et al.*, *Nat. Plants* **9**, 572–587 (2023).
16. P. J. Wittkopp, G. Kalay, *Nat. Rev. Genet.* **13**, 59–69 (2012).
17. A. P. Marand, A. L. Eveland, K. Kaufmann, N. M. Springer, *Annu. Rev. Plant Biol.* **74**, 111–137 (2023).
18. H. K. Long, S. L. Prescott, J. Wysocka, *Cell* **167**, 1170–1187 (2016).
19. F. W. Albert, L. Kruglyak, *Nat. Rev. Genet.* **16**, 197–212 (2015).
20. W. Schwarzer, F. Spitz, *Curr. Opin. Genet. Dev.* **27**, 74–82 (2014).
21. S. Kim, J. Wysocka, *Mol. Cell* **83**, 373–392 (2023).
22. T. Fuqua *et al.*, *Nature* **587**, 235–239 (2020).
23. J. H. Massey, J. Li, D. L. Stern, P. J. Wittkopp, *Heredity* **127**, 467–474 (2021).
24. M. Kitagawa, D. Jackson, *Annu. Rev. Plant Biol.* **70**, 269–291 (2019).
25. H. Schoof *et al.*, *Cell* **100**, 635–644 (2000).
26. D. Rodríguez-Leal, Z. H. Lemmon, J. Man, M. E. Bartlett, Z. B. Lippman, *Cell* **171**, 470–480.e8 (2017).
27. D. Rodriguez-Leal *et al.*, *Nat. Genet.* **51**, 786–792 (2019).
28. X. Wang *et al.*, *Nat. Plants* **7**, 419–427 (2021).
29. C.-T. Kwon *et al.*, *Nat. Plants* **8**, 346–355 (2022).
30. G. C. Briggs, K. S. Osmont, C. Shindo, R. Sibout, C. S. Hardtke, *Trends Plant Sci.* **11**, 492–498 (2006).
31. L. Pereira *et al.*, *Theor. Appl. Genet.* **134**, 3363–3378 (2021).
32. M. Alonge *et al.*, *Cell* **182**, 145–161.e23 (2020).
33. Y. Zhou *et al.*, *Nature* **606**, 527–534 (2022).
34. T. F. Hansen, G. P. Wagner, *Theor. Popul. Biol.* **59**, 61–86 (2001).
35. C. W. Bakerlee, A. N. Nguyen Ba, Y. Shulgina, J. I. Rojas Echenique, M. M. Desai, *Science* **376**, 630–635 (2022).
36. D. M. Lyons, Z. Zou, H. Xu, J. Zhang, *Nat. Ecol. Evol.* **4**, 1685–1693 (2020).
37. A. B. Paaby, M. V. Rockman, *Nat. Rev. Genet.* **15**, 247–258 (2014).
38. K. McGuigan, N. Nishimura, M. Currey, D. Hurwit, W. A. Cresko, *Evolution* **65**, 1203–1211 (2011).
39. J. Hou, J. van Leeuwen, B. J. Andrews, C. Boone, *Trends Genet.* **34**, 578–586 (2018).
40. N. Lauter, J. Doebley, *Genetics* **160**, 333–342 (2002).
41. M. Fagny, F. Austerlitz, *Trends Genet.* **37**, 631–638 (2021).
42. M. Costanzo *et al.*, *Cell* **177**, 85–100 (2019).
43. R. Kafri, M. Springer, Y. Pilpel, *Cell* **136**, 389–392 (2009).
44. G. Diss, D. Ascencio, A. DeLuna, C. R. Landry, *J. Exp. Zool. B Mol. Dev. Evol.* **322**, 488–499 (2014).
45. F. J. Yuste-Lisbona *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 8187–8195 (2020).
46. X. Song *et al.*, *Nat. Biotechnol.* **40**, 1403–1411 (2022).
47. P. E. Bayer, A. A. Golicz, A. Scheben, J. Batley, D. Edwards, *Nat. Plants* **6**, 914–920 (2020).
48. R. M. Sherman, S. L. Salzberg, *Nat. Rev. Genet.* **21**, 243–254 (2020).
49. J. A. Draghi, J. B. Plotkin, *Evolution* **67**, 3120–3131 (2013).
50. J. Doebley, A. Stec, C. Gustus, *Genetics* **141**, 333–346 (1995).
51. L. Aguirre, A. Hendelman, S. F. Hutton, D. M. McCandlish, Z. B. Lippman, davidmccandlish/Aguirre-etal-epistasis: Data for: Idiosyncratic and dose-dependent epistasis drives variation in tomato fruit size, Zenodo (2023); https://doi.org/10.5281/zenodo.8267283.

ACKNOWLEDGMENTS

We thank members of the Lippman laboratory for comments, discussions, and assistance with phenotyping. We thank M. Bartlett and Y. Eshed for helpful discussions. We thank B. Seman and G. Robitaille from the Lippman lab for technical support. We thank T. Mulligan, K. Schlecht, A. Krainer, S. Qiao, and B. Fitzgerald for assistance with plant care. **Funding:** This research was funded by NSF Graduate Research Fellowship grant 1938105 (L.A.), William Randolph Hearst Foundation Scholarship (L.A.), NIH grant R35GM133613 (D.M.M.), Alfred P. Sloan Research Fellowship (D.M.M.), NSF Plant Genome Research Program grant IOS-2129189

(Z.B.L.), and the Howard Hughes Medical Institute (HHMI) (Z.B.L.)
Author contributions: Conceptualization: D.M.M and Z.B.L. Methodology: L.A., S.F.H., D.M.M., and Z.B.L. Investigation: L.A., S.F.H., A.H., D.M.M., and Z.B.L. Visualization: L.A., A.H., D.M.M., and Z.B.L. Funding acquisition: L.A., D.M.M., and Z.B.L. Supervision: Z.B.L. Writing – original draft: L.A., A.H., D.M.M., and Z.B.L. Writing – review and editing: L.A., A.H., D.M.M., and Z.B.L. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Source code for statistical analysis of epistasis models can be found on Zenodo (*51*). All data are available in the main text or the supplementary materials. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse. This article is subject to HHMI's Open Access to Publications policy. HHMI lab heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript (AAM) of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adi5222
Materials and Methods
Figs. S1 to S3
Tables S1 to S4
Reference (*52*)
MDAR Reproducibility Checklist

Submitted 5 May 2023; accepted 6 September 2023
10.1126/science.adi5222

FOOD WEB ECOLOGY

Periodical cicadas disrupt trophic dynamics through community-level shifts in avian foraging

Zoe L. Getman-Pickering[1]†, Grace J. Soltis[2]‡, Sarah Shamash[1], Daniel S. Gruner[2], Martha R. Weiss[3]*§, John T. Lill[1]*§

Once every 13 or 17 years within eastern North American deciduous forests, billions of periodical cicadas concurrently emerge from the soil and briefly satiate a diverse array of naive consumers, offering a rare opportunity to assess the cascading impacts of an ecosystem-wide resource pulse on a complex food web. We quantified the effects of the 2021 Brood X emergence and report that more than 80 bird species opportunistically switched their foraging to include cicadas, releasing herbivorous insects from predation and essentially doubling both caterpillar densities and accumulated herbivory levels on host oak trees. These short-lived but massive emergence events help us to understand how resource pulses can rewire interaction webs and disrupt energy flows in ecosystems, with potentially long-lasting effects.

Regional biomass pulses, such as the spawning of salmon or the masting of oak trees, can provide enormous nutritional subsidies to entire communities of predators, producers, and decomposers that capitalize on these intermittent resources (*1*–*4*). Biomass pulses can also alter species interactions, initiating cascades of direct and indirect effects (*5*–*7*). For example, when generalist predators bypass their typical prey items in favor of an abundant pulsed resource, the organisms experiencing reduced predation benefit indirectly—a form of "apparent facilitation" (Fig. 1) (*8*). Such pulse-induced dietary shifts have the potential to disrupt top-down trophic cascades, which occur when the foraging activities of predators benefit producers through reductions in herbivore densities (*9*). By altering the flow of energy through ecosystems, biomass pulses of sufficient extent and magnitude are thus likely to produce landscape-level ecological impacts.

Insectivorous birds, which are common drivers of trophic cascades, are ubiquitous predators in most terrestrial ecosystems. As adaptive foragers that respond functionally and numerically to abundant resources in time and space (*10*), birds provide an important ecosystem service by suppressing both perennial insect herbivore communities and outbreaks of herbivorous insect pests (*11*). The resultant reduction in herbivory can allow for increased plant biomass, growth, and/or yield in both natural and managed ecosystems (*12*). Researchers have estimated that insectivorous birds routinely provide billions of dollars in pest-reduction services to agricultural systems through top-down control (*11*). Given the inherent challenges of studying trophic cascades at landscape scales, regional perturbations—such as those resulting from the emergences of periodical cicadas—provide natural experiments that allow researchers to explore how biomass pulses alter species interactions and disrupt the functioning of food webs.

Periodical cicada (*Magicicada* spp.) (Fig. 2) emergence events in the eastern United States are among the largest and most dramatic biomass pulses found in nature. In the spring of an emergence year, billions of cicada nymphs synchronously crawl up and out of their underground burrows to molt into winged adults. Males congregate in trees, chorusing loudly to attract females. Once mated, individual females deposit up to 500 eggs in twigs and die shortly after. Later in the summer, 2-mm-long nymphs hatch, drop to the soil, and burrow underground, where they will develop for the next 13 or 17 years, feeding on xylem sap of host tree roots (*13*). The seven described species of *Magicicada* are distributed among 15 geographically distinct broods with differing emergence schedules. Each brood spans an area of 50,000 to 500,000 km^2 (*14*), and most contain three to four sympatric *Magicicada* species that share an emergence year. Despite centuries of detailed investigation into the biology, behavior, and evolution of these intriguing insects (*13*), researchers have only recently begun to explore the ecological consequences of periodical cicada emergences on forest ecosystems.

We report the results of a community-level study in which we examined how a periodical cicada emergence affected avian foraging behaviors and documented the consequences of these behavioral shifts for local trophic dynamics. We first tested the hypothesis that avian consumers would capitalize on the abundant, nutritious cicadas, as shown previously after outbreaks of other native and invasive insects (*10*, *15*). We further predicted that shifts in avian foraging to cicadas would reduce predation pressure on foliage-feeding insects, consequently increasing insect herbivore populations and herbivory on their host trees. Therefore, in both emergence and non-emergence years, we (i) quantified avian predation on caterpillars using plasticine models (*16*), (ii) quantified understory insect herbivore abundance, and (iii) measured seasonal patterns of herbivory on understory white oaks (*Quercus alba*) (Fig. 1).

Cicada emergence causes widespread changes in avian foraging

In addition to our intensive field observations of avian foraging during the 2021 Brood X emergence, we solicited foraging observations from volunteer birders throughout the region. Altogether, we recorded 983 individual birds from 82 different species foraging on cicadas throughout the Mid-Atlantic portion of the Brood X range. These birds spanned more than three orders of magnitude in body mass, ranging from <7 g (blue-gray gnatcatchers) to >11,000 g (trumpeter swans), and included all major avian feeding guilds (fig. S1 and tables S1 to S3) (*17*). Focal observations revealed that 52% of all avian foraging bouts (n = 229) involved periodical cicadas. Unexpectedly, the proportion of foraging bouts that included cicadas did not vary among feeding guilds (df = 2, χ^2 = 3.6, P = 0.31) or with bird size (df = 1, χ^2 = 0.65, P = 0.42) (*17*).

The collective response of the avian community to the cicada pulse documented in this study extended well beyond that of the subset of birds observed to have altered their foraging during outbreaks of spruce budworm (*10*) or autumnal moths (*18*). Several factors likely contributed to this community-wide dietary shift, beyond the sheer abundance of nutritious, undefended prey. These factors include the availability of cicadas in a wide array of avian foraging niches, ranging from bare ground to canopy foliage, and the temporal overlap between the emergence event and the energetically demanding breeding season (*19*).

Because adult periodical cicadas are relatively large insects (3 to 4 cm in length), we had hypothesized that their consumption might be limited by avian gape width (*20*) or bird size more generally, as has been reported previously (*21*, *22*). However, we observed varied prey-processing behaviors that allowed birds of all sizes to feed successfully on cicadas (*17*). The similar likelihood of cicada consumption by birds across feeding guilds illustrates the birds' ability to respond to pulses through opportunistic diet shifts, regardless of normal dietary preferences.

[1]Department of Biological Sciences, George Washington University, Washington, DC 20052, USA. [2]Department of Entomology, University of Maryland, College Park, MD 20742, USA. [3]Department of Biology, Georgetown University, Washington, DC 20007, USA.
*Corresponding author. Email: weissm@georgetown.edu (M.R.W.); lillj@gwu.edu (J.T.L.)
†Present address: College of Engineering, University of Massachusetts, Amherst, MA 01003, USA.
‡Present address: Department of Biology, Florida State University, Tallahassee, FL 32306, USA.
§These authors contributed equally to this work.

Cicada emergence reduces avian predation on forest caterpillars

As anticipated, heavy consumption of cicadas by a diverse assemblage of birds at our field site resulted in a measurable decline in bird predation on oak-feeding caterpillars, which in a typical year are consumed in large numbers by insectivorous birds (*23*). Relative to the flanking nonemergence years, the weekly frequency of bird strikes on plasticine caterpillars (Fig. 3A) dipped dramatically during the 2021 season [generalized linear mixed model (GLMM), df = 24, χ^2 = 102, P < 0.001]. Specifically, in May, when the cicada nymphs began emerging, bird strike frequency declined from ~30% to below 10%, remaining low until early July (Fig. 3B), after which we ceased to hear adult males calling at our study site. By early August, bird strike frequencies had returned to preemergence levels. By contrast, in nonemergence years, the weekly mean frequency of avian predation never dropped below ~25% (Fig. 3B), despite some expected seasonal variation. In 2021, we concurrently monitored a second study site and recorded equally reduced levels of bird strikes during the emergence (estimated marginal means, z = −0.978, P = 0.33). Both the within- and between-year patterns of predation pressure indicate a strong avian response to cicada presence.

To assess the consequences of the cicada emergence for real insect herbivores, we conducted nondestructive, whole-tree censuses of externally feeding foliage feeders, consisting primarily of caterpillars (n = 26 species) and sawflies (n = 6 species) (table S4) (*17*), on a cohort of understory white oak trees in early May (timed to precede the cicada pulse in 2021) and late June of 2021, 2022, and 2023. These trees hosted similar herbivore abundances in early May of 2021 and 2022, but had reduced numbers in 2023 [generalized linear model (GLM), χ^2 = 50.14, df = 2, P < 0.0001)] (Fig. 4A).

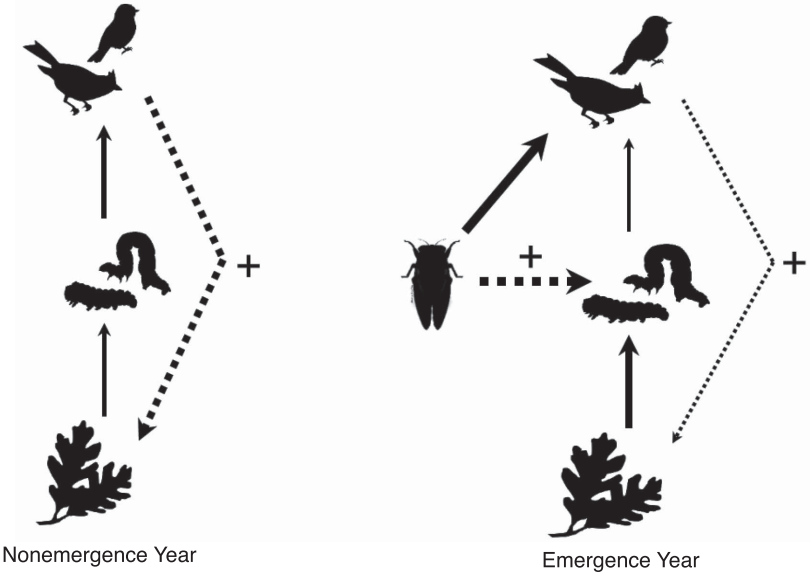


Fig. 1. Predicted impacts of cicada emergence on trophic dynamics. Heuristic food web models illustrating how multitrophic interactions are hypothesized to change from (**Left**) nonemergence years to (**Right**) cicada emergence years. Solid arrows indicate direction of energy flow through direct trophic effects, dashed arrows indicate positive indirect effects, and line thickness indicates relative effect magnitudes. In nonemergence years, bird consumption of caterpillars initiates a top-down trophic cascade that results in a strong positive indirect effect on oaks. In emergence years, we predicted that bird consumption of abundant cicadas would reduce bird consumption of caterpillars, weakening the positive indirect effect of birds on oaks, but yielding a strong positive indirect effect of cicadas on caterpillars through apparent facilitation.

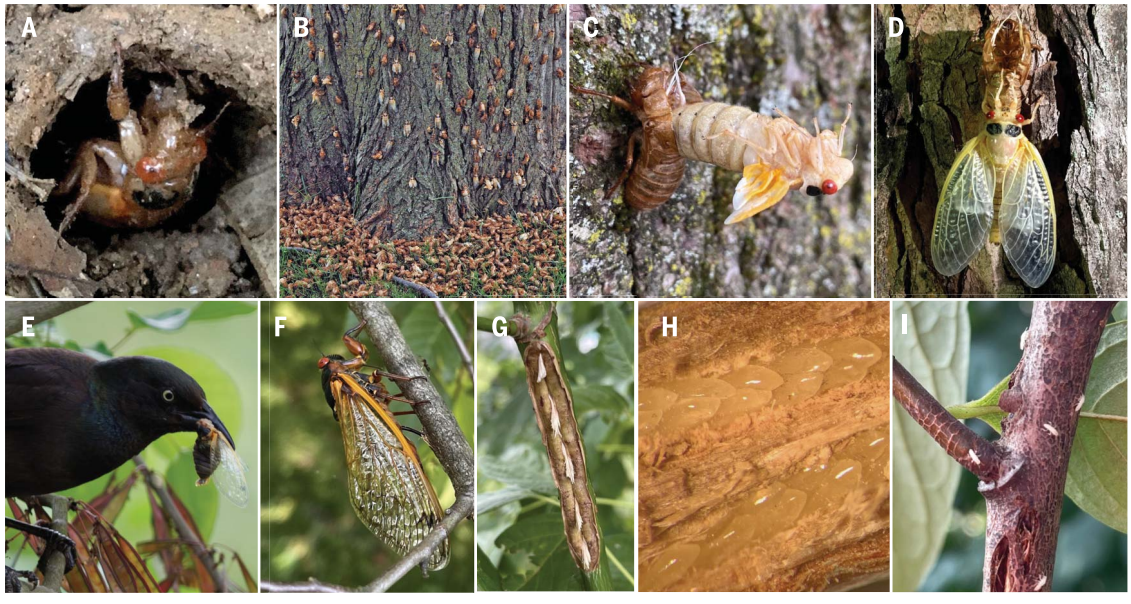


Fig. 2. Life cycle of periodical *Magicicada* spp. Three species of Brood X periodical cicadas (*Magicicada septendecim*, *M. cassini*, and *M. septendecula*) emerged by the billions across 15 US states in May and June of 2021 to begin the brief above-ground phase of their 17-year life cycle. (**A**) Nymphs emerge from underground burrows, after spending 17 years feeding on the xylem fluids of tree roots. (**B**) Massive numbers of insects emerge over a short period of time, in densities reported to be as high as >300/m^2. (**C**) Shrimp-like nymphs soon molt into winged adults. (**D**) Cuticles of the soft, light-colored teneral-phase adults harden and darken over the course of 2 to 3 days. (**E**) More than 80 species of birds, including this common grackle (*Quiscalus quiscula*), were observed consuming cicada nymphs or adults. (**F**) After mating, females deposit multiple batches of approximately 30 eggs into nests slit into slender woody branches, with each female laying a total of up to 500 eggs. (**G**) Females often lay several egg nests in a row. (**H**) Eggs develop in the nests for about 6 weeks. (**I**) In midsummer, tiny hatchlings drop from the branches and burrow into the soil, where they locate and begin to feed from a root and will remain for the next 17 years. [Photo credits: (A), (C), (D), (F), (G), and (I), M. Weiss; (B), M. Raupp; (E), D. Gruner; (H), E. K. Lederer].

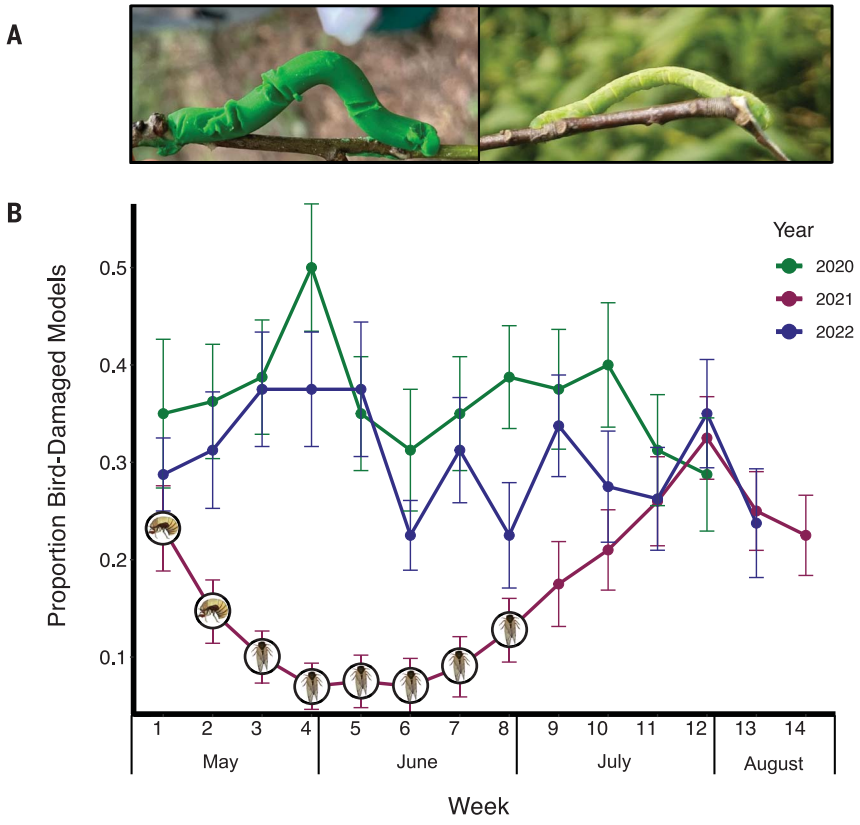


Fig. 3. Weekly measures of avian predation pressure (2020 to 2022). (**A**) Plasticine models (5 cm long) mimicking naturally occurring forest caterpillars were deployed on understory white oak trees and replaced weekly. (**B**) Plot of the average (±1 SEM) proportion of plasticine models with bird strikes recorded throughout the spring and summer of the emergence (2021) and flanking nonemergence (2020 and 2022) years. Symbols for the 2021 data indicate the presence of nymph and adult cicadas.

Notably, during the late June census of the emergence year, the abundance of externally feeding insect herbivores was essentially double that of each of the two subsequent nonemergence years (GLM χ^2 = 142.59, df = 2, P < 0.0001) (Fig. 4A). For *Acronicta increta*, one of the most common caterpillar species at our site (Fig. 4B), the significant spike in abundance in the emergence year (GLM χ^2 = 94.83, df = 2, P < 0.0001) (Fig. 4C) was accompanied by a marked shift in larval demography: By recording the size of each individual *A. increta* caterpillar present on the trees (n = 421 total, across the 3 years) as "small" (first or second larval instars) or "large" (third to fifth larval instars), we were able to document a fourfold to 50-fold increase in the proportion of large caterpillars present in the emergence year (52% in 2021) relative to the subsequent two nonemergence years (13% in 2022 and <1% in 2023) (χ^2 = 101.7, df = 2, P < 0.0001) (Fig. 4B). This demographic shift suggests that the normally high predation pressure experienced by larger, more conspicuous *A. increta* caterpillars was temporarily relaxed during the 2021 cicada emergence.

Taken together, these findings support the hypothesis that a community-wide avian diet shift was responsible for the dramatic release in predation we observed in the understory throughout the emergence. The pronounced dip in bird strikes on plasticine models during the emergence and the subsequent return to baseline levels was observed at two different Maryland sites separated by ~100 km, suggesting that this foraging shift was widespread. Because adult cicadas congregate along forest edges (*24*), it is likely that birds shift their foraging locations to track cicada population densities. Although tracking an entire bird community was beyond the scope of this project, a previous study documented that red-winged blackbirds (*Ageleius phoeniceus*) adaptively switched from feeding on aquatic insects in open marshes to foraging for periodical cicadas in forest edges (*25*). A similar locational shift may have contributed to the increase in understory herbivore abundance reported here.

Cicada emergence indirectly increases oak herbivory

At the producer level, understory white oak trees experienced a spike in cumulative leaf damage in midsummer 2021 [analysis of variance (ANOVA), with Tukey's post-hoc comparisons: $F_{2,87}$ = 12.3, P < 0.001] relative to both the preemergence (t = −4.66, df = 87, P < 0.0001) and postemergence years (t = 3.789, df = 87, P < 0.0001), which did not differ from one another (t = −0.88, df = 87, P = 0.66) (Fig. 5). This increase in herbivory coincided with the observed peak in caterpillar density recorded during the emergence period, indicating that the community-wide shift in avian foraging cascaded down to affect plant damage (Fig. 1).

Although increased herbivory on understory white oaks has been documented previously when birds were excluded from individual trees (*23*) or small stands (*26*), we provide the first empirical evidence that the protective function of an entire community of insectivorous birds can be disrupted by a synchronous resource pulse. Because we studied prereproductive white oak trees, we could not assess the fitness effects of herbivory in this study, but prior research on mature white oak trees has demonstrated that comparable levels of damage reduce acorn production (*26*–*28*).

Ecological significance

The enormous biomass pulse created by the 2021 Brood X emergence shifted consumption by mobile generalists (birds) to superabundant prey resources (cicadas), releasing insect herbivore populations from predation and doubling herbivory on oak trees. This event constitutes a clear example of apparent facilitation (*8*), in which the brief appearance of a single brood of *Magicicada* spp. provides a month-long predation respite for a diverse community of leaf-chewing insects across a broad geographical region (for example, up to 500,000 km^2 for Brood X) (*29*). Our empirical findings support the view (*30*) that insect herbivore communities are more often structured by indirect interactions [such as apparent competition, mutualism, or facilitation (*31*, *32*)] than by direct competition, and our findings add to the accumulating body of research stressing the importance of considering positive interactions in community ecology (*33*, *34*).

Our study demonstrates that periodical cicada emergences can "rewire" forest food webs (*35*, *36*), altering interaction strengths and pathways of energy flow that affect multiple trophic levels. Adaptive foraging, in which mobile consumers converge on elevated resources, is widely observed in response to resource pulses (*37*, *38*). By focusing consumption on hyperabundant resources while releasing less-common resources, adaptive foraging can promote complexity in food webs, stabilize their dynamics, and alter the strength of indirect effects such as trophic cascades (*39*, *40*).

Although previous studies have documented strong ecological impacts that result from the direct consumption of pulsed resources (*21*, *41*), our results highlight the need to consider the

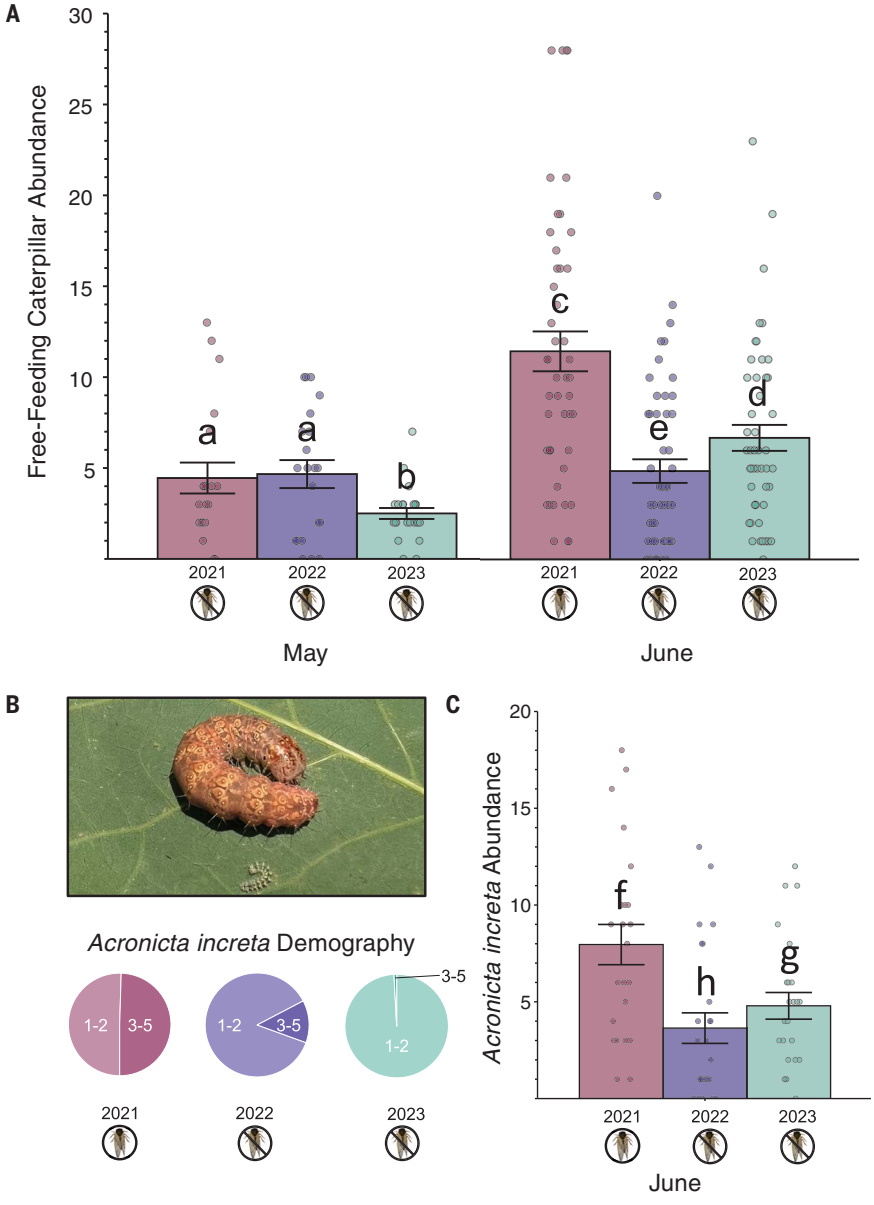


Fig. 4. Abundance of oak herbivores (2021 to 2023). (**A**) Model-estimated marginal mean abundances (±1 SEM) of externally feeding insect herbivores recorded on understory white oaks. Trees were censused in May (just before adult periodical cicada emergence) and June (toward the end of the emergence) during the emergence year (2021) and two subsequent nonemergence years (2022 and 2023). (**B**) Small (first instar) and large (fifth instar) *A. increta* caterpillars differ dramatically in apparency. Pie graphs depict the proportion of *A. increta* caterpillars categorized as small (first or second larval instars) or large (third to fifth larval instars). Large *A. increta* caterpillars were 4 to >50 times more abundant during the emergence year (52%) than in the nonemergence years (13 and <1%, respectively). (**C**) Model-estimated marginal mean abundances (±1 SEM) of the most common externally feeding caterpillar species (*A. increta*) on the same trees in emergence and nonemergence years. For (A) and (C), raw data are overplotted and jittered; one high outlier was removed for clarity. Means with the same letter are not significantly different according to Tukey-adjusted post-hoc comparisons of marginal means within a grouping.

additional effects stemming from the organisms that escape predation during pulses. As a classic example of the former, regional oak masting events in North American forests can increase populations of acorn-consuming small mammals that serve as the primary reservoirs of the spirochaete *Borrelia burgdorferi* and as larval hosts of black-legged ticks, leading to increased risk of Lyme disease for humans (*42*). However, if the consumption of seeds of co-occurring trees (such as beeches and hickories) decreases when oaks are masting, seedling recruitment may increase for those species, potentially affecting forest structure. Broadly, whenever pulsed resources satiate generalist predators, we predict that community-wide trophic disruptions will likely follow, and ecologists should examine the trophic impacts that result both from what is consumed (the pulsed resource) and what is not consumed (the typical prey) during these pulse events.

Although periodical cicada emergences are short-lived phenomena, lasting for only 5 to 6 weeks, they can initiate a cascade of ecological impacts that propagate up and down the food chain. By providing abundant living prey to predators, carcasses to scavengers and decomposers, and an infusion of nutrients into both terrestrial (*2*) and aquatic (*43*) ecosystems, these brief yet intense events initiate a wide range of responses that can last from days to years and can occur immediately or with a considerable time lag. For instance, analyses of long-term data collected by the North American Breeding Bird Survey revealed that populations of some species were abundant only during emergence years and subsequently declined (for example, black- and yellow-billed cuckoos), whereas others were scarce during emergence years but increased substantially the following year, and then stabilized (for example, tufted titmice and gray catbirds) (*21*). Migratory birds exhibiting cicada-mediated numerical responses can potentially extend the cicadas' legacy across continents when they travel to distant overwintering sites, affecting recipient communities through "cross-boundary subsidy cascades" (*7*, *44*).

At the producer level, the nutrient addition from decomposing cicadas can alter foliar nitrogen content, nutrient use efficiency, plant biomass, and patterns of herbivory in herbaceous plants (*6*, *45*), indicating complex bottom-up plant response to these pulses. Dendrochronological analyses have demonstrated that radial growth in oak trees is predictably suppressed in cicada emergence years (*46*), potentially because of the extensive twig damage that occurs during oviposition (*21*, *47*). Recent research has shown that periodical cicada emergences also set the timing of oak masting events, linking two of the largest and most pervasive resource pulses in eastern temperate forests and greatly extending the temporal, taxonomic, and trophic reach of these emergence events (*47*). In addition, cicada emergences also leave long-lasting imprints on the physical environment; the billions of persistent underground burrows and emergence tunnels constructed by cicada nymphs dramatically increase soil aeration and water infiltration (*48*), a type of physical ecosystem engineering (*49*) that is likely to affect plants and other soil-dwelling organisms.

In any given spring, an emergence of one or more of the 15 extant broods of *Magicicada*

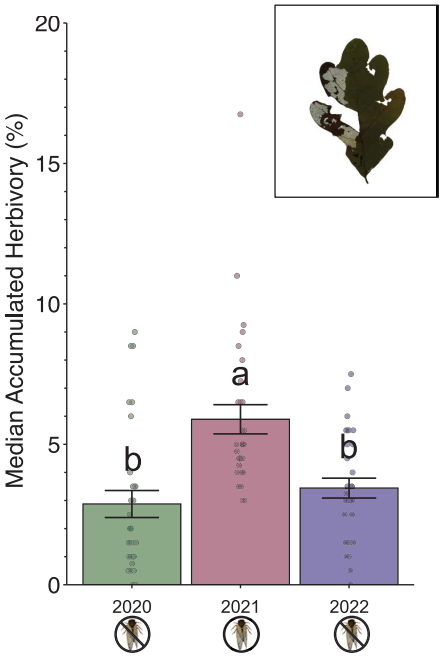


Fig. 5. Herbivory on understory oaks (2020 to 2022). Mean (±1 SEM) accumulated herbivory on white oak trees during the summer months of each of 3 years (after the spring pulse of herbivory, which occurs before the cicada emergence). Accumulated herbivory was measured as the difference in median herbivory calculated from samples of 30 leaves per tree in late May and mid-October of each year, allowing us to capture the herbivory during the cicada emergence time period in both emergence and nonemergence years. Raw data are jittered, with each point indicating accumulated herbivory measured on an individual tree. Means with the same letter are not significantly different according to Tukey-adjusted post-hoc comparisons of marginal means.

spp. is imminent, in progress, or has recently taken place somewhere in the eastern deciduous forest biome of North America (*50*). Our study suggests that these insects leave a legacy of ecological impacts at the landscape scale, altering the dynamics of the populations, communities, and ecosystems within which they are embedded.

REFERENCES AND NOTES

1. R. D. Holt, *Ecology* **89**, 671–681 (2008).
2. L. H. Yang, *Science* **306**, 1565–1567 (2004).
3. L. H. Yang, J. L. Bastow, K. O. Spence, A. N. Wright, *Ecology* **89**, 621–634 (2008).
4. L. H. Yang *et al*., *Ecol. Monogr.* **80**, 125–151 (2010).
5. P. Chesson *et al*., *Oecologia* **141**, 236–253 (2004).
6. L. H. Yang, *Arthropod-Plant Interact.* **7**, 93–98 (2013).
7. R. Zwolak, J. Witczuk, M. Bogdziewicz, L. Rychlik, S. Pagacz, *J. Mammal.* **99**, 586–595 (2018).
8. C. E. Wilkinson *et al*., *Conserv. Biol.* **34**, 854–867 (2020).
9. J. Terborgh, J. A. Estes, Eds., *Trophic Cascades: Predators, Prey, and the Changing Dynamics of Nature* (Island Press, 2010).
10. H. S. Crawford, D. T. Jennings, *Ecology* **70**, 152–163 (1989).
11. C. J. Whelan, D. G. Wenny, R. J. Marquis, *Ann. N. Y. Acad. Sci.* **1134**, 25–60 (2008).
12. K. A. Mooney *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **107**, 7335–7340 (2010).
13. C. Simon, J. R. Cooley, R. Karban, T. Sota, *Annu. Rev. Entomol.* **67**, 457–482 (2022).
14. G. Kritsky, *Periodical Cicadas: The Brood X Edition* (Ohio Biological Survey, 2021).
15. R. T. Holmes, T. W. Sherry, F. W. Sturges, *Ecol. Monogr.* **56**, 201–220 (1986).
16. P. W. Bateman, P. A. Fleming, A. K. Wolfe, *J. Zool.* **301**, 251–262 (2017).
17. Materials and methods are available as supplementary materials.
18. O. Hogstad, *Ibis* **147**, 77–91 (2005).
19. K. S. Williams, K. G. Smith, F. M. Stephen, *Ecology* **74**, 1143–1152 (1993).
20. N. T. Wheelwright, *Ecology* **66**, 808–818 (1985).
21. W. D. Koenig, A. M. Liebhold, *Ecology* **86**, 1873–1882 (2005).
22. K. Tomita, *Naturwissenschaften* **108**, 52 (2021).
23. R. J. Marquis, C. J. Whelan, *Ecology* **75**, 2007–2014 (1994).
24. K. S. Williams, C. Simon, *Annu. Rev. Entomol.* **40**, 269–295 (1995).
25. C. E. Strehl, J. White, *Oecologia* **70**, 178–186 (1986).
26. M. Murakami, S. Nakano, *Proc. Biol. Sci.* **267**, 1597–1601 (2000).
27. M. J. Crawley, *Nature* **314**, 163–164 (1985).
28. C. G. Hochwender, V. L. Sork, R. J. Marquis, *Am. Midl. Nat.* **150**, 246–253 (2003).
29. J. C. Blackwood *et al*., *Am. Nat.* **192**, 479–489 (2018).
30. I. Kaplan, R. F. Denno, *Ecol. Lett.* **10**, 977–994 (2007).
31. R. D. Holt, M. B. Bonsall, *Annu. Rev. Ecol. Evol. Syst.* **48**, 447–471 (2017).
32. W. C. Long, E. F. Gamelin, E. G. Johnson, A. H. Hines, *Mar. Ecol. Prog. Ser.* **456**, 139–148 (2012).
33. A. A. Agrawal *et al*., *Front. Ecol. Environ.* **5**, 145–152 (2007).
34. M. D. Bertness, R. Callaway, *Trends Ecol. Evol.* **9**, 191–193 (1994).
35. F. Polazzo, T. I. Marina, M. Crettaz-Minaglia, A. Rico, *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2117364119 (2022).
36. T. J. Bartley *et al*., *Nat. Ecol. Evol.* **3**, 345–354 (2019).
37. J. B. Armstrong, G. Takimoto, D. E. Schindler, M. M. Hayes, M. J. Kauffman, *Ecology* **97**, 1099–1112 (2016).
38. H. V. Kenny *et al*., *Ecol. Evol.* **7**, 10701–10709 (2017).
39. F. S. Valdovinos, R. Ramos-Jiliberto, L. Garay-Narváez, P. Urbani, J. A. Dunne, *Ecol. Lett.* **13**, 1546–1559 (2010).
40. A. K. Fahimipour, K. E. Anderson, *Ecol. Lett.* **18**, 826–833 (2015).
41. D. T. Krohne, T. J. Couillard, J. C. Riddle, *Am. Midl. Nat.* **126**, 317–321 (1991).
42. R. S. Ostfeld, C. D. Canham, K. Oggenfuss, R. J. Winchcombe, F. Keesing, *PLOS Biol.* **4**, e145 (2006).
43. C. L. Pray, W. H. Nowlin, M. J. Vanni, *J. N. Am. Benthol. Soc.* **28**, 181–195 (2009).
44. M. S. Luskin *et al*., *Nat. Commun.* **8**, 2231 (2017).
45. L. H. Yang, *Ecology* **89**, 1497–1502 (2008).
46. R. Karban, *Nature* **287**, 326–327 (1980).
47. W. D. Koenig, A. M. Liebhold, J. M. LaMontagne, I. S. Pearse, *Am. Nat.* **201**, 755–762 (2023).
48. D. L. Ficklin *et al*., *Hydrol. Processes* **37**, e14822 (2023).
49. C. G. Jones, J. H. Lawton, M. Shachak, *Ecology* **78**, 1946–1957 (1997).
50. G. Kritsky, *Am. Entomol.* **67**, 40–46 (2021).
51. Z. L. Getman-Pickering *et al*., *Dryad* (2023); doi:10.5061/dryad.n5tb2rc1s.
52. Z. L. Getman-Pickering *et al*., *Zenodo* (2023); doi:10.5281/zenodo.8237533.

ACKNOWLEDGMENTS

We thank E. Lederer, R. Spahn, D. Ertem, B. Thexton, C. Wallsten, S. Hermelin, L. Micheels, I. Rosenthal, F. Grifo, and C. Robarts for providing invaluable field and laboratory assistance; J. Parker and K. Burghardt for facilitating research at SERC; P. Marra and D. Narango for assistance in the design and analysis of bird foraging data; the many volunteer birders who generously contributed bird foraging data to this project, especially D. Levey for his numerous contributions; the Bethesda-Chevy Chase Chapter of the IWLA for providing access to our primary field site; and R. Marquis, K. Gedan, C. Hochwender, D. Lill, J. Rosenthal, A. Getman-Pickering, and members of the insect ecology laboratory groups at GWU, GU, and UMD who provided critical feedback to earlier versions of this manuscript. **Funding:** This work was supported by the George Washington University Facilitating Fund (J.T.L.); The Earth Commons: Georgetown University's Institute for Environment and Sustainability (M.R.W. and J.T.L.); and Washington Biologists' Field Club research grants (J.T.L., M.R.W., Z.G.P.). **Author contributions:** Conceptualization: Z.L.G.-P., M.R.W., and J.T.L. Funding acquisition and project supervision: Z.L.G.-P., M.R.W., and J.T.L. Field Data collection: Z.L.G.-P., G.J.S., S.S., J.T.L., and M.R.W. Data entry: Z.G.P., S.S., G.S., and J.T.L. Data analysis and figure and table preparation: Z.G.P., M.R.W., J.T.L., G.S., and D.S.G. Writing – original draft: M.R.W. and J.T.L. Writing – review and editing: Z.G.P., S.S., G.S., D.S.G., M.R.W., and J.T.L. **Competing Interests:** The authors declare that they have no competing interests. **Data and materials availability:** Data and code have been archived in Dryad (*51*) and zenodo (*52*), respectively. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adi7426
Materials and Methods
Supplementary Text
Fig. S1
Tables S1 to S4
References (*53*–*66*)
MDAR Reproducibility Checklist

Submitted 15 May 2023; accepted 25 August 2023
10.1126/science.adi7426

ORGANIC CHEMISTRY

Taming secondary benzylic cations in catalytic asymmetric S_N1 reactions

Vikas Kumar Singh[1]†, Chendan Zhu[1]†, Chandra Kanta De[1]†, Markus Leutzsch[1], Lorenzo Baldinelli[2], Raja Mitra[3], Giovanni Bistoni[2], Benjamin List[1]*

Benzylic stereogenic centers are ubiquitous in natural products and pharmaceuticals. A potentially general, though challenging, approach toward their selective creation would be asymmetric unimolecular nucleophilic substitution (S_N1) reactions that proceed through highly reactive benzylic cations. We now report a broadly applicable solution to this problem by identifying chiral counteranions that pair with secondary benzylic cations to engage in catalytic asymmetric C–C, C–O, and C–N bond-forming reactions with excellent enantioselectivity. The critical cationic intermediate can be accessed from different precursors via Lewis- or Brønsted acid catalysis. Key to our strategy is the use of only weakly basic, confined counteranions that are posited to prolong the lifetime of the carbocation, thereby avoiding nonproductive deprotonation pathways to the corresponding styrene.

Benzylic stereocenters are omnipresent substructures of a vast number of natural products and drugs (*1*). For the asymmetric construction of the crucial benzylic C–C or C–heteroatom bond, a plethora of methods has been described. Approaches that use chiral auxiliaries, metal catalysts, organocatalysts, and enzymes, among others, are known (*2*–*5*). However, despite its potential as a general approach to benzylic stereocenters, the catalytic enantioselective reaction of a nucleophile with a benzylic cation remains underdeveloped. During the past decade, early attempts toward this strategy have been reported but typically involved oxygenated arenes, which stabilize the corresponding benzylic cation as an *ortho*- or *para*-quinone methide, thus limiting the scope of these methods (*6*–*12*).

Recently, the Jacobsen group reported an allylation of tertiary benzylic cations for the construction of quaternary stereogenic centers using anion-binding, hydrogen bond–donor catalysis (Fig. 1A) (*13*). Subsequently, the Sun group explored tertiary benzylic cations as intermediates in asymmetric catalysis (*14*). The asymmetric catalysis of reactions that proceed through unstabilized secondary benzylic cations has previously been reported by the Braun and Toste groups, albeit each with only a single example (Fig. 1A) (*15*, *16*). We hypothesized that engaging unbiased secondary benzylic carbocationic intermediates with a wide range of nucleophiles could provide a general solution to the problem of constructing benzylic stereocenters (Fig. 1B).

Over the past several years, highly acidic and enzyme-like confined imidodiphosphorimidates (IDPi) catalysts have emerged as a powerful motif to control enantioselectivity in the reactions of high-energy, cationic intermediates such as tertiary benzylic carbocations, vinyl carbocations, and the nonclassical 2-norbornyl cation (*17*–*20*). Initially, we expected three fundamental challenges toward applying our design to unstabilized secondary benzylic cations: (i) the differentiation between two faces of an only hydrocarbon-based planar cationic intermediate, (ii) the exclusion of unproductive reaction pathways such as deprotonation or rearrangement, and (iii) potential catalyst deactivation by alkylation from the benzylic cation. Given the high local electrophilicity and hydride ion affinities of benzylic cations, we postulated that essential cation stabilization could be achieved using a less basic, weakly coordinating anion, which, at the same time, can also provide a chiral microenvironment for further nucleophilic attack to achieve high enantiocontrol (*21*–*23*). In this work, we report the design and development of an IDPi-catalyzed S_N1 platform that enables the conversion of racemic sp^3 starting materials into valuable enantioenriched benzylic stereocenter–containing products in a dynamic kinetic asymmetric transformation.

Reaction development

First, we explored C–O bond-forming reactions between *rac*-1-phenylethyl 2,2,2-trichloroacetimidate (**1a**) as the cation precursor and acetic

[1]Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, D-45470 Mülheim an der Ruhr, Germany. [2]Department of Chemistry, Biology and Biotechnology, University of Perugia, 06122 Perugia, Italy. [3]School of Chemical and Materials Sciences, Indian Institute of Technology Goa, Farmagudi, Ponda, Goa-403401, India.
*Corresponding author. Email: list@kofo.mpg.de
†These authors contributed equally to this work.

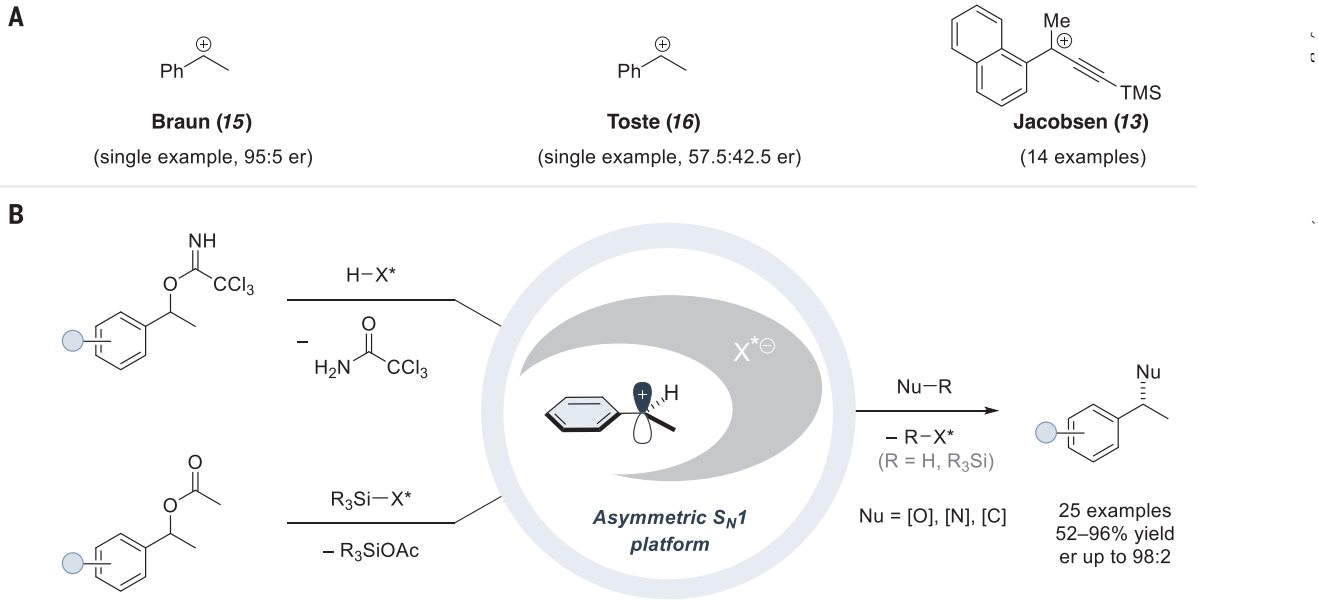


Fig. 1. Subjecting benzylic cations to asymmetric catalysis. (**A**) Dynamic kinetic asymmetric transformations via benzylic cations. (**B**) This work: Confined counteranions tame secondary benzylic cations for asymmetric catalysis. Ac, acetyl; Me, methyl; Nu, nucleophile; Ph, phenyl; TMS, trimethylsilyl; X*, enantiopure counteranion.

acid as the nucleophile in the presence of different IDPi catalysts. In our initial attempts, we found that an IDPi catalyst (**S2h**) could catalyze this reaction, giving the desired product **3a** in 86% yield with a 61:39 enantiomeric ratio (er), so further optimizations were carried out (see supplementary materials for details). Eventually, IDPi **2a** was identified as the optimal catalyst, delivering product **3a** in 73% yield and with a 95:5 er (Fig. 2A). With the optimal conditions (IDPi **2a**, temperature $T = -90°C$, time $t = 5$ days) established, we turned our attention toward exploring the scope for this transformation. Substrates (**1b** to **1d**) with meta or para substituents performed well and furnished the products with high enantioselectivities and modest to good yields. When propionic acid was used as the nucleophile, fluorenyl substituted IDPi catalyst **2b** demonstrated superior performance. Propionate products **3e** to **3h** were obtained in good yields and with high enantioselectivities.

These results are consistent with a facile ionization of trichloroacetimidate **1a**, even at low temperature, delivering the highly reactive benzylic carbocation along with trichloroacetamide. We wondered whether the in situ–generated trichloroacetamide could itself act as an *N*-nucleophile in the absence of an external *O*-nucleophile. Such a reaction might serve as an attractive approach to the enantioselective formation of benzylic C–N bonds from benzylic alcohols (*24*–*28*). However, trichloroacetamide is only poorly soluble in diethyl ether at low temperature, and an undesired deprotonation to styrene could be problematic. Indeed, upon testing various IDPi catalysts, the desired rearrangement product **4a** was obtained with moderate to good enantioselectivities but in only poor yield with substantial amounts of styrene as a side product (see supplementary materials for details). We reasoned that the undesired deprotonation pathway could be minimized by using even less-basic IDPi counteranions, which would require more-acidic IDPi catalysts. As shown in our previous work, electron-withdrawing groups on the 1,1′-bi-2-naphthol (BINOL) backbone enhance the acidity of IDPi catalysts (*29*). Therefore, we attempted to diminish the deprotonation pathway by modulating the electronic properties of the catalyst. Gratifyingly, F_8-substitution of one BINOL skeleton led to catalyst **2c**, which provided the desired product **4a** in 75% yield and with an excellent 95:5 er (Fig. 2B). Moreover, substrates **1b** to **1e**, which bear alkyl groups and halogen atoms at the meta or para positions, provided the corresponding products with good to excellent enantioselectivities and moderate to good yields using catalyst **2d**.

Extension to carbon nucleophiles

After the successful implementation of these Brønsted acid–catalyzed asymmetric C–O and C–N bond-forming S_N1 reactions, we were keen on exploring the exciting potential to create C–C bonds by a similar pathway. We envisioned the previously unknown reactions of secondary benzylic alcohol derivatives with silyl ketene acetals or electron-rich arenes toward the formation of β-branched esters or 1,1-diaryl ethanes, respectively, as particularly attractive targets. A somewhat related copper-catalyzed enantioselective substitution of benzylic propargylic acetates using indoles has been described (*30*). However, with silyl ketene acetal, only 10% yield and no enantioinduction was obtained. Similarly, our initial attempts at reacting trichloroacetimidate **1a** with silyl ketene acetals (SKAs) using IDPi catalysts did not lead to the desired C–C bond-forming product, and only the substrate **1a** was recovered. We speculate that imidate **1a** engages in a deprotosilylation (*31*), the product of which, however, is unreactive

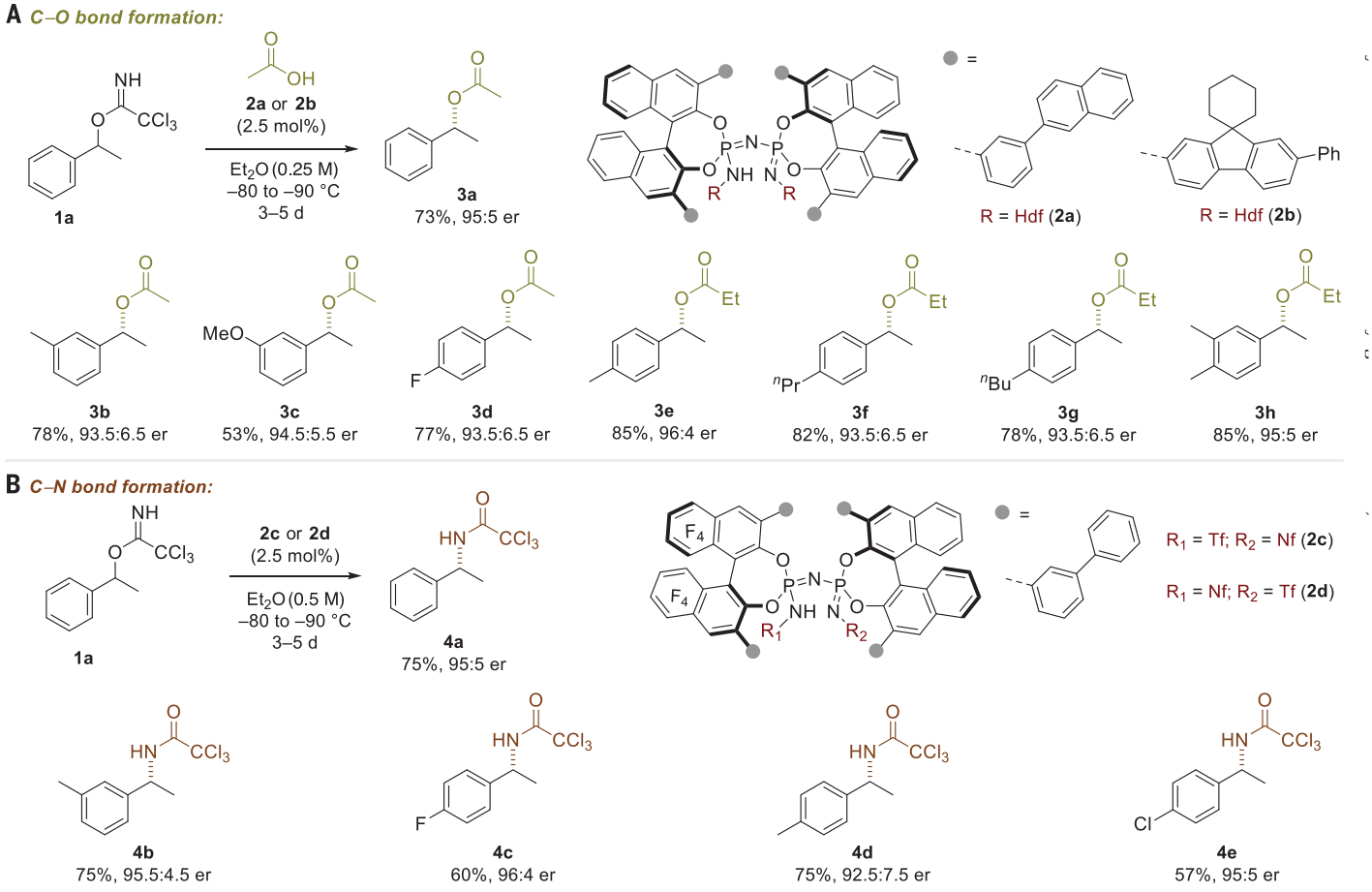


Fig. 2. Substrate scope of the Brønsted acid–catalyzed C–O and C–N bond-forming reactions. (**A**) C–O bond-forming reaction. (**B**) C–N bond-forming reaction. See supplementary materials for details. ^{n}Bu, *n*-butyl; Et, ethyl; Et_2O, diethyl ether; Hdf, $-SO_2C_8F_{17}$; Nf, $-SO_2C_4F_9$; ^{n}Pr, *n*-propyl; Tf, SO_2CF_3.

toward ionization to the reactive benzylic carbocation. We had previously shown that heteroatom-stabilized cations (oxocarbenium ions and iminium ions) could be generated under (oxophilic) silylium Lewis acid catalysis conditions when acetate was used as leaving group (*32*). We therefore investigated racemic 1-phenylethyl acetate (**5a**) as substrate toward a silylium Lewis acid–catalyzed cation formation. Indeed when acetate **5a** was treated with {[1-(benzyloxy)vinyl]oxy}(tert-butyl)dimethylsilane (**8a**) in the presence of IDPi catalysts, the desired product **6a** was obtained in good yields and with moderate to good enantioselectivities (see supplementary materials for details). Upon optimization, we found that IDPi catalyst **2e** provided product **6a** with a 92.5:7.5 er and in 85% yield (Fig. 3A). Other substrates with alkyl substituents at the para position, linear or branched, and with different chain lengths delivered products **6b** to **6f** in 84 to 91% yields with high enantioselectivities using fluorenyl-substituted IDPi catalyst **2f**. Substrate **5g**, which has meta, para disubstitution, furnished product **6g** in 92% yield with 95:5 er. Furthermore, product **6h** was obtained with modest enantioselectivity using catalyst **2e**, where a methoxy group in place of acetate was used as the leaving group for substrate **5h** because the corresponding acetate substrate was found to be unstable for isolation. The microenvironment of the binding site in the IDPi catalyst allows C–C bond formation to proceed by selective addition of a single SKA nucleophile. Presumably, the steric differentiation between substrate and product prevents the product from participating further in a Mukaiyama-Claisen–type condensation reaction.

Asymmetric Friedel-Crafts alkylations

With the successful establishment of a silylium Lewis acid–catalyzed C–C-bond-forming α-alkylation reaction, we were keen to expand the asymmetric transformation of secondary benzylic cations with other C-nucleophiles. The Friedel-Crafts alkylation as a classic and broadly applicable transformation that proceeds via carbocations became a natural target. Previously, asymmetric catalytic Friedel-Crafts alkylations that involve benzylic cations had not been reported. Compared with carboxylic acids and SKAs, electron-rich arenes are much weaker nucleophiles (*33*, *34*). We therefore envisioned controlling and stabilizing the benzylic cation in this electrophilic aromatic substitution reaction a rather challenging task. Gratifyingly, under silylium Lewis acid conditions, the Friedel-Crafts alkylation of racemic acetate substrate **5a** with anisole gave product **7a** in 60% yield with a 96:4 er and >50:1 regioisomeric ratio (rr) using IDPi catalyst **2g** (Fig. 3B). Different anisoles were evaluated under these conditions with the same catalyst. When anisoles with ortho and meta substituents were applied in this transformation, products **7b** and **7c** were obtained with excellent enantioselectivities, regioselectivity, and modest yields. Notably, a silylated phenol was found to be a suitable nucleophile for this transformation and delivered the trimethylsilylated product **7d** in good yield with high enantioselectivity. Furthermore, this conversion also proceeded under purely Brønsted acid catalytic conditions using substrate **1a** and the same IDPi catalyst **2g** to furnish the same product with consistent enantioselectivity, in line with the formation of a benzylic cation via an S_N1-like mechanism (see supplementary materials for details).

Mechanistic considerations

Experimental and computational studies were carried out to challenge our benzylic cation S_N1 hypothesis (Fig. 4A). At first, substrate *rac*-**1a** and each of the enantiopure starting materials,

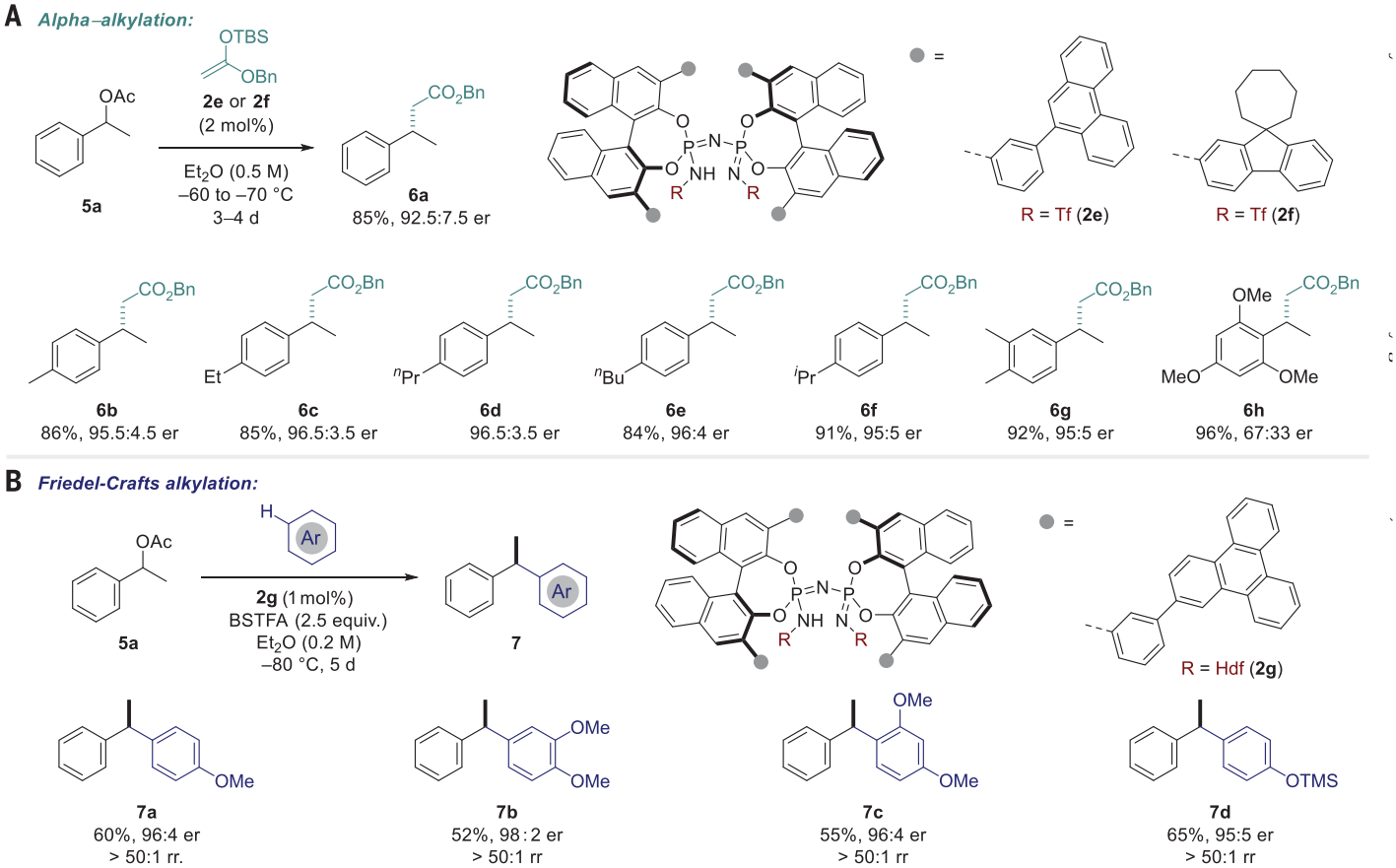


Fig. 3. Substrate scope of the silylium Lewis acid–catalyzed C–C bond-forming reactions. (**A**) SKA α-alkylation reaction. (**B**) Friedel-Crafts reaction. See supplementary materials for details. Bn, benzyl; BSTFA, *N*,*O*-bis(trimethylsilyl)trifluoroacetamide; OMe, methoxy; OTMS, trimethylsilyloxy; OTBS, tert-butyldimethylsilyloxy.

(*S*)-**1a** or (*R*)-**1a**, were individually subjected to the optimized reaction conditions. Consistent with ion-pair formation followed by nucleophilic attack on the benzylic cation in a S_N1-like mechanism, we found that all three reactions delivered the same major enantiomer of product **3a** with essentially identical yield [85, 86, and 85% by nuclear magnetic resonance (NMR)] and enantioselectivity (95:5, 95:5, and 94.5:5.5 er) (Fig. 4B). Interestingly, when the progress of the C–O bond-forming reaction was monitored by ^{1}H NMR at −60°C for each substrate separately [*rac*-**1a**, enantiopure (*S*)-**1a**, and (*R*)-**1a**], both enantiomers were found to react with similar rates, showing very little to no kinetic resolution (Fig. 4C). Attempts to characterize either an ion pair or a covalent adduct in the reaction of substrate **1a** with IDPi **2a** were unsuccessful, exclusively delivering product **4a** and styrene under the reaction conditions. To gain further insight into the cationic intermediates and their lifetimes, we investigated a series of biased substrates featuring different substitution patterns, which were ionized with bis(trifluoromethanesulfonyl)amide ($HNTf_2$) as a strong achiral acid. In most cases, the ion pair was either not detectable or rapidly underwent decomposition during the measurement (see supplementary materials for details). We anticipated that the ion pair that resulted from the reaction of 1-(2,4,6-trimethoxyphenyl)ethan-1-ol (**5i**) and $HNTf_2$ would have a sufficiently long lifetime for its characterization. Indeed, when alcohol **5i** was treated with $HNTf_2$ (3 equiv) in CD_2Cl_2 at −60°C in an NMR tube, the formation of the ion pair was observed and could be characterized by NMR spectroscopy (^{1}H, ^{13}C, and 2D NMR experiments) (Fig. 4, D and E). A gradual increase of the temperature from −60° to −20°C was performed, and the resulting mixture was monitored by NMR spectroscopy. Ion pair **9** was found to be stable in that range and to only begin decomposing at higher temperature. The same ion pair **9** could also be generated under silylium Lewis acid conditions, and a similar

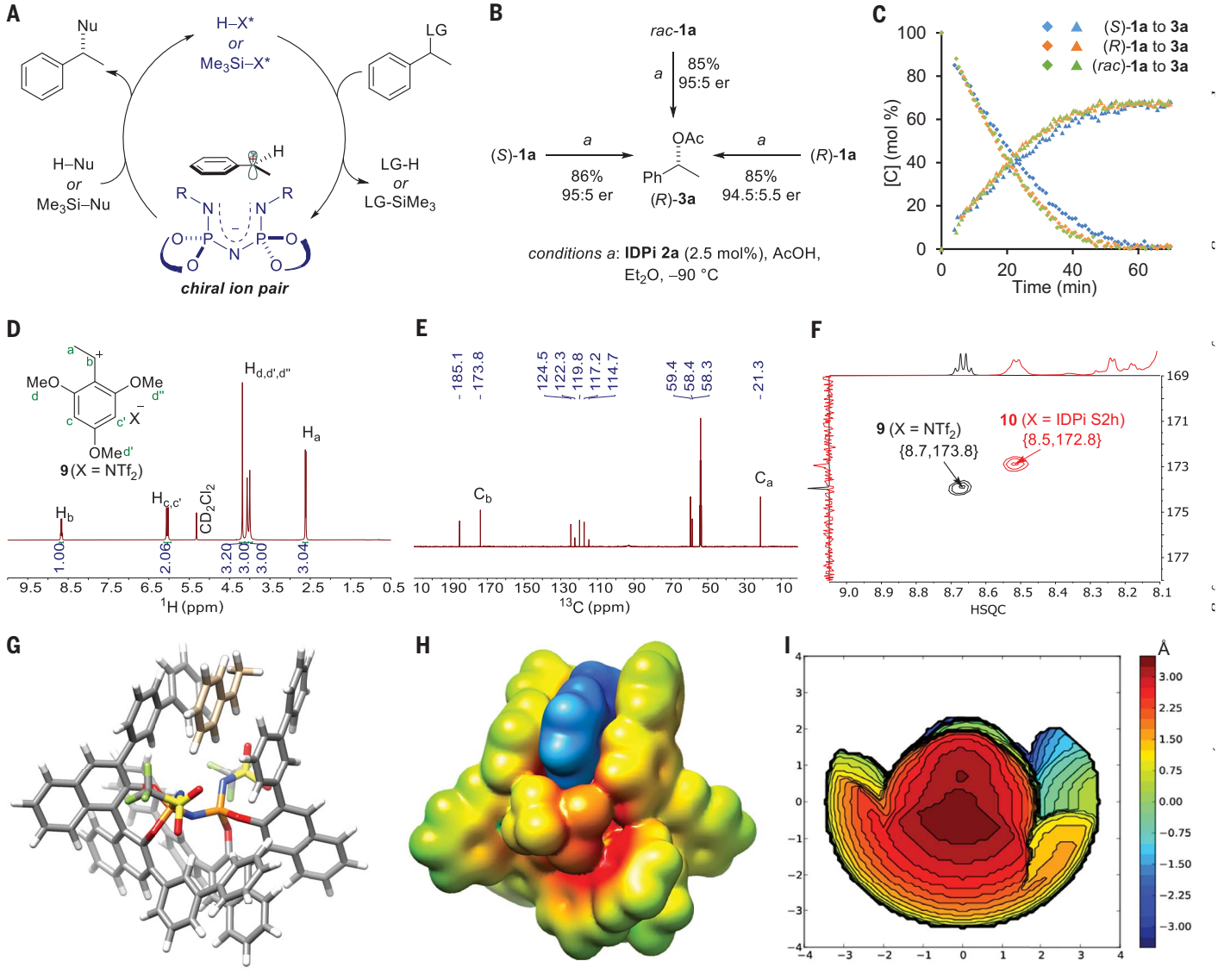


Fig. 4. Mechanistic studies, characterization of the ion pair, and origin of enantioselectivity. (**A**) Proposed reaction pathway. (**B**) Enantioselectivity and NMR yield comparison of *rac*-**1a** with enantiopure **1a** substrate. (**C**) Reactivity measurements of *rac*-**1a** and enantiopure **1a**. The reaction was monitored by ^{1}H NMR at −60°C. (**D**) ^{1}H NMR spectrum of an ion pair (**9**) consisting of 1-(2,4,6-trimethoxyphenyl)ethan-1-ylium cation and bistriflimide anion. (**E**) ^{13}C NMR spectrum of the ion pair (**9**). (**F**) Comparison of heteronuclear single quantum coherence (HSQC) spectrum for bistriflimide and IDPi (**S2h**) anions. (**G**) DFT-optimized ion-pair structure. (**H**) Electrostatic potential map (MEP) of the ion pair. (**I**) Topographic steric map of the ion pair (see supplementary materials for orientation and further details). AcOH, acetic acid; LG, leaving group.

behavior was observed in this case (see supplementary materials for details). Importantly, the oxygenated benzylic cation could also be observed with catalyst **S2h** in NMR experiments. Decomposition of this ion-pair species **10** was observed at −60°C, presumably because of the relatively higher basicity of the IDPi counteranion compared with that of the NTf_2 anion. A shift in the ^{1}H and ^{13}C NMR [benzylic proton from 8.7 to 8.5 parts per million (ppm) and benzylic carbon from 173.8 to 172.8 ppm] was observed (Fig. 4F), clearly indicating different counteranions in close proximity to the cation. In addition, density functional theory (DFT) was used to optimize the ion pair. A large number of conformers was initially generated at various levels of GFNn-FF/XTB theory, and the low-energy structures were refined using DFT (see supplementary materials for details). Analysis of the electrostatic potential surfaces revealed that the ion-pair structure is held together by directional electrostatic interactions between the IDPi anion and the benzylic carbocation (Fig. 4, G to I). The computed association free energy between the IDPi anion and benzylic carbocation (ΔG) is negative (−22.08 kcal/mol), indicating that the formation of the ion pair is thermodynamically favored. The decomposition of the electronic interaction energy between the anion and the cation (ΔE_{int} = −43.21 kcal/mol) into dispersive (ΔE_{disp}) and steric and electrostatic components ($\Delta E_{ster/elec}$) shows that dispersion plays a crucial role in favoring the formation of the chiral ion pair, contributing 27.33 kcal/mol to the ΔE_{int}. The negative interaction is at least partially counterbalanced by (i) the energy penalty required to distort the catalyst to maximize the electrostatic interaction with the cation and (ii) entropy and temperature effects. Most notably, the cation resides between the 3-biphenyl substituents of the BINOL skeleton (3.56 and 3.19 Å, respectively) that aid in stabilizing the reactive intermediate by virtue of cation-π interactions. Additionally, the benzylic CH group clearly interacts with a sulfonyl oxygen atom (2.21 Å), and the phenyl 3-CH interacts with an inner-core nitrogen atom (2.31 Å; for a full display of interactions, see fig. S6). It is plausible to assume that all of these interactions can be expected to play a vital role in the stabilization of the respective transition states leading to the major enantiomer. Given the importance of carbocations in chemical synthesis, our results may have broader implications for a plethora of other fundamental organic transformations.

REFERENCES AND NOTES

1. N. A. McGrath, M. Brichacek, J. T. Njardarson, *J. Chem. Educ.* **87**, 1348–1349 (2010).
2. G. C. Fu, *ACS Cent. Sci.* **3**, 692–700 (2017).
3. D. T. Ziegler, G. C. Fu, *J. Am. Chem. Soc.* **138**, 12069–12072 (2016).
4. S. Athavale, K. Chen, F. H. Arnold, in *Transition Metal–Catalyzed Carbene Transformations*, J. Wang, C.-M. Che, M. P. Doyle, Eds. (Wiley, 2022), chap. 4.
5. D. A. Evans, D. Ennis, D. J. Mathre, *J. Am. Chem. Soc.* **104**, 1737–1739 (1982).
6. Z. Wang, J. Sun, *Synthesis* **47**, 3629–3644 (2015).
7. A. A. Jaworski, K. A. Scheidt, *J. Org. Chem.* **81**, 10145–10153 (2016).
8. W. Li, X. Xu, P. Zhang, P. Li, *Chem. Asian J.* **13**, 2350–2359 (2018).
9. J. Kikuchi, M. Terada, *Chemistry* **27**, 10215–10225 (2021).
10. C. Dorsch, C. Schneider, *Synthesis* **54**, 3125–3141 (2022).
11. X. Li, Z. Li, J. Sun, *Nat. Synth.* **1**, 426–438 (2022).
12. J. Ouyang, R. Maji, M. Leutzsch, B. Mitschke, B. List, *J. Am. Chem. Soc.* **144**, 8460–8466 (2022).
13. A. E. Wendlandt, P. Vangal, E. N. Jacobsen, *Nature* **556**, 447–451 (2018).
14. X. Tan *et al.*, *Nat. Syn.* **2**, 275–285 (2023).
15. M. Braun, W. Kotter, *Angew. Chem. Int. Ed.* **43**, 514–517 (2004).
16. C. Zhao, F. D. Toste, K. N. Raymond, R. G. Bergman, *J. Am. Chem. Soc.* **136**, 14409–14412 (2014).
17. N. Tsuji *et al.*, *Science* **359**, 1501–1505 (2018).
18. P. Zhang, N. Tsuji, J. Ouyang, B. List, *J. Am. Chem. Soc.* **143**, 675–680 (2021).
19. R. Properzi *et al.*, *Nat. Chem.* **12**, 1174–1179 (2020).
20. S. K. Nistanaki *et al.*, *Science* **378**, 1085–1091 (2022).
21. A. Aizman, R. Contreras, P. Pérez, *Tetrahedron* **61**, 889–895 (2005).
22. D. H. Aue, in *Dicoordinated Carbocations*, Z. Rappoport, P. J. Stang, Eds. (Wiley, 1997), pp. 105–156.
23. E. V. Anslyn, D. A. Dougherty, *Modern Physical Organic Chemistry* (University Science, 2006), p. 88.
24. Z. Li, Z. Zhou, L. Wang, Q. Zhou, C. Tang, *Tetrahedron Asymmetry* **13**, 145–148 (2002).
25. K. Tauber *et al.*, *Chemistry* **19**, 4030–4035 (2013).
26. F. G. Mutti, T. Knaus, N. S. Scrutton, M. Breuer, N. J. Turner, *Science* **349**, 1525–1529 (2015).
27. F.-F. Chen, Y.-Y. Liu, G.-W. Zheng, J.-H. Xu, *ChemCatChem* **7**, 3838–3841 (2015).
28. E. Liardo, N. Ríos-Lombardía, F. Morís, F. Rebolledo, J. González-Sabín, *ACS Catal.* **7**, 4768–4774 (2017).
29. O. Grossmann, R. Maji, M. H. Aukland, S. Lee, B. List, *Angew. Chem. Int. Ed.* **61**, e202115036 (2022).
30. R. J. Detz, Z. Abiri, R. le Griel, H. Hiemstra, J. H. van Maarseveen, *Chemistry* **17**, 5921–5930 (2011).
31. H. Zhou *et al.*, *J. Am. Chem. Soc.* **142**, 13695–13700 (2020).
32. S. Lee, P. S. J. Kaib, B. List, *J. Am. Chem. Soc.* **139**, 2156–2159 (2017).
33. H. Mayr, B. Kempf, A. R. Ofial, *Acc. Chem. Res.* **36**, 66–77 (2003).
34. H. F. Schaller, A. A. Tishkov, X. Feng, H. Mayr, *J. Am. Chem. Soc.* **130**, 3012–3022 (2008).

ACKNOWLEDGMENTS

We thank the technicians of our group and the members of our gas chromatography, mass spectrometry, high-performance liquid chromatography, and NMR service departments. We also thank O. Grossmann, V. A. F. de Silva, and B. Mitschke for their help and members of our group for in-house reviewing. **Funding:** We acknowledge generous support from the Max Planck Society, the Deutsche Forschungsgemeinschaft Cluster of Excellence [Ruhr Explores Solvation (RESOLV)], and the European Research Council [Early Stage Organocatalysis (ESO)] to B.L. **Author contributions:** Conceptualization: B.L.; Methodology: V.K.S., C.Z., C.K.D., R.M.; Investigation: V.K.S., C.Z., C.K.D., R.M.; Project administration: B.L.; Supervision: B.L.; NMR: M.L.; DFT: G.B., L.B.; Writing – original draft: V.K.S., C.Z., C.K.D., B.L.; Writing – review and editing: V.K.S., C.Z., C.K.D., B.L. **Competing interests:** A patent WO2017037141 (A1) filed by the Max-Planck-Institut für Kohlenforschung and B.L. covers the IDPi catalyst class and its applications in asymmetric synthesis. The authors declare no other competing interests. **Data and materials availability:** All data and synthetic details used in the analysis are available in supporting material. **License information:** Copyright © 2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adj7007
Materials and Methods
Supplementary Text
Figs. S1 to S8
Tables S1 to S17
NMR Spectra
HPLC and GC Traces
References (*35*–*54*)

Submitted 11 July 2023; accepted 5 September 2023
10.1126/science.adj7007

COMPUTER SCIENCE

Neural inference at the frontier of energy, space, and time

Dharmendra S. Modha*, Filipp Akopyan†, Alexander Andreopoulos†, Rathinakumar Appuswamy†, John V. Arthur†, Andrew S. Cassidy†, Pallab Datta†, Michael V. DeBole†, Steven K. Esser†, Carlos Ortega Otero†, Jun Sawada†, Brian Taba†, Arnon Amir, Deepika Bablani, Peter J. Carlson, Myron D. Flickner, Rajamohan Gandhasri, Guillaume J. Garreau, Megumi Ito, Jennifer L. Klamo, Jeffrey A. Kusnitz, Nathaniel J. McClatchey, Jeffrey L. McKinstry, Yutaka Nakamura, Tapan K. Nayak, William P. Risk, Kai Schleupen, Ben Shaw, Jay Sivagnaname, Daniel F. Smith, Ignacio Terrizzano, Takanori Ueda

Computing, since its inception, has been processor-centric, with memory separated from compute. Inspired by the organic brain and optimized for inorganic silicon, NorthPole is a neural inference architecture that blurs this boundary by eliminating off-chip memory, intertwining compute with memory on-chip, and appearing externally as an active memory chip. NorthPole is a low-precision, massively parallel, densely interconnected, energy-efficient, and spatial computing architecture with a co-optimized, high-utilization programming model. On the ResNet50 benchmark image classification network, relative to a graphics processing unit (GPU) that uses a comparable 12-nanometer technology process, NorthPole achieves a 25 times higher energy metric of frames per second (FPS) per watt, a 5 times higher space metric of FPS per transistor, and a 22 times lower time metric of latency. Similar results are reported for the Yolo-v4 detection network. NorthPole outperforms all prevalent architectures, even those that use more-advanced technology processes.

In the design of the first programmable computer, EDVAC (*1*), compute and memory used heterogeneous technologies (vacuum tubes versus ultrasonic serial delays) that had disparate relative speeds and costs, leading to the separation of compute from memory by necessity and, eventually, to a hierarchy of memories. These fundamental constructs still pervade modern computing where energy and bandwidth for off-chip memory dominate (*2*–*4*), creating a "data movement crisis" (*5*) and forming a physical and an "intellectual" programming model (*6*) bottleneck. Three factors necessitate revisiting these constructs for neural inference. First, inspired by the brain (*7*), neural inference (*8*, *9*) has emerged as a powerful application that can be implemented with a far simpler set of constructs. Second, silicon technology is continuing to progress such that, within a decade, 2 billion transistors may fit in 1 mm^2 (*10*). In the meantime, logic and memory have long been implemented homogeneously within the same substrate (*11*, *12*), although at a lower density and higher cost than off-chip memory. Third, it is now critical to prevent artificial intelligence (AI) from becoming "economically, technically, and environmentally unsustainable" through "dramatically more computationally-efficient methods" (*13*).

Axiomatic design

NorthPole, an architecture and a programming model for neural inference, reimagines (Fig. 1) the interaction between compute and memory by embodying 10 interrelated, synergistic axioms that build on brain-inspired computing (*14*, *15*). Within the broad domain of neural inference, NorthPole co-optimizes architecture, algorithms, and software and enables a simple input-output model (*16*).

To set terminology, a neural network is typically a graph of layers, where each layer transforms an input frame of neural activations (for example, an image) into an output frame of neural activations (for example, a class label). Typically, for each neuron, this involves multiplications of its input neuron activations with learned synaptic layer weights followed by linear accumulation, mathematical transforms, and possibly a nonlinear activation function. The synaptic multiplications are inherently parallel for every neuron, and all neurons within a layer operate in parallel.

Axiom 1: Turning to architecture, NorthPole is specialized for neural inference. For example, it has no data-dependent conditional branching, and it does not support training or scientific computation.

Axiom 2: Inspired by biological precision (*17*), NorthPole is optimized for 8-, 4-, and 2-bit low-precision (*5*). This is sufficient to achieve state-of-the-art inference accuracy on many neural networks while dispensing with the high-precision required for training.

Axiom 3: NorthPole has a distributed, modular core array (16-by-16), with each core capable of massive parallelism (8192 2-bit operations per cycle) (Fig. 2F). Cortex-like modularity (*18*) of the tiled core array enables homogeneous scalability in two dimensions and, perhaps, even in three dimensions (*19*) and is also amenable to heterogeneous chiplet integration.

Axiom 4: NorthPole distributes memory among cores (Figs. 1B and 2F) and, within a core, not only places memories near compute (*2*) but also intertwines critical compute with memory (Fig. 2, A and B). The nearness of memory and compute enables each core to exploit locality for energy efficiency. NorthPole dedicates a large area to on-chip memory that is neither centralized nor organized in a traditional memory hierarchy.

Axiom 5: NorthPole uses two dense networks-on-chip (NoCs) (*20*) to interconnect the cores, unifying and integrating the distributed computation and memory (Fig. 2, C and D) that would otherwise be fragmented. These NoCs are inspired by long-distance white-matter and short-distance gray-matter pathways in the brain and by neuroanatomical topological maps found in cortical sensory and motor systems (*21*). One gray matter–inspired NoC enables spatial computing between adjacent cores (Fig. 3 and fig. S1). Another white matter–inspired NoC enables neuron activations to be spatially redistributed among all cores.

Axiom 6: Another two NoCs enable reconfiguring synaptic weights and programs on each core for high-speed operation of compute units (Fig. 2, C and D). The brain's organic biochemical substrate is suitable for supporting many slow analog neurons, where each neuron is hardwired to a fixed set of synaptic weights. Directly following this architectural construct leads to an inefficient use of inorganic silicon, which is suitable for fewer and faster digital neurons. Reconfigurability resolves this key dilemma by storing weights and programs just once in the distributed memory and reconfiguring the weights during the execution of each layer using one NoC and reconfiguring the programs before the start of the layer using another NoC. Stated differently, these two NoCs serve to substantially increase (up to 256 times, in some cases) the effective on-core memory sizes for weights and programs such that each core computes as if the weights and program for the entire network are stored on every core. Consequently, NorthPole achieves 3000 times more computation and 640 times larger network models than TrueNorth (*14*), although it has only four times more transistors (supplementary text S1).

Axiom 7: NorthPole exploits data-independent branching to support a fully pipelined, stall-free, deterministic control operation for high temporal utilization without any memory misses, which are a hallmark of the von Neumann architecture. Lack of memory misses eliminates the need for speculative, nondeterministic execution. Deterministic operation enables a set of eight threads for various compute, memory, and communication operations to be synchronized by construction and to operate at a high utilization.

Axiom 8: Turning to algorithms and software, co-optimized training algorithms (fig. S3) enable state-of-the-art inference accuracy to be achieved by incorporating low-precision constraints into training. Judiciously selecting precision for each layer enables optimal use of on-chip resources

IBM Research, San Jose, CA, USA.
*Corresponding author. Email: dmodha@us.ibm.com
†These authors contributed equally to this work.

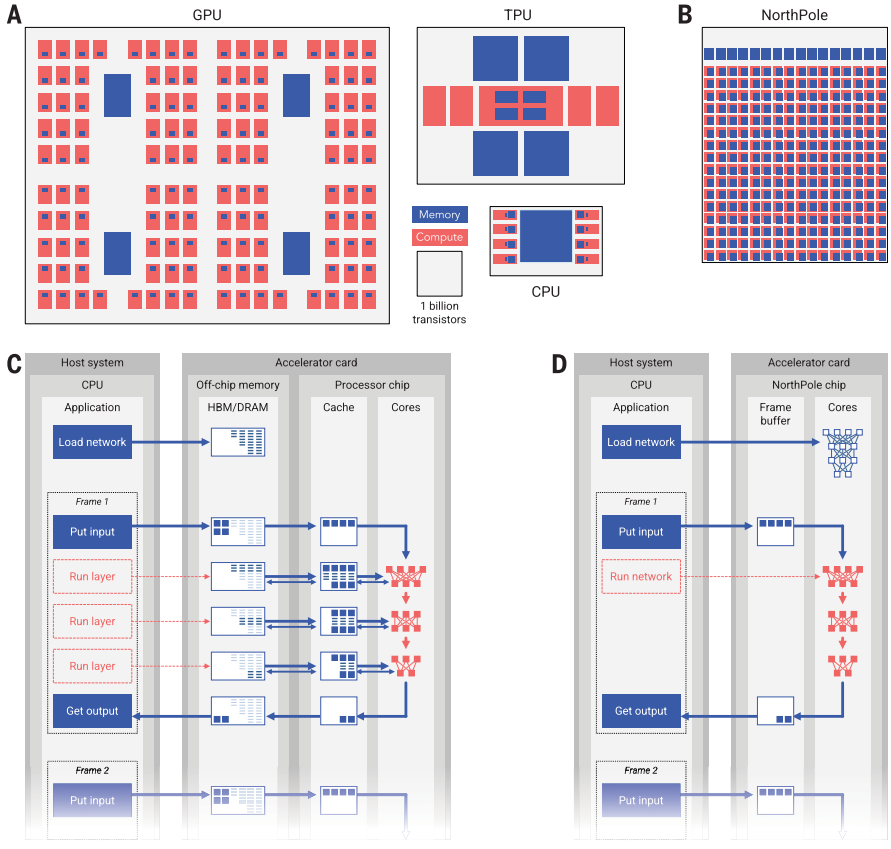


Fig. 1. Memory near compute and active memory: Processor memory organization dictates on-chip data pathways and usage model. (**A**) Physical organization of on-chip memory (blue) and compute (red) are diagrammed for representative processors, scaled to constant transistors per unit area. Large central blocks of memories separate compute from memory in central, tensor, and graphics processing units (CPU, TPU, and GPU)—which use off-chip memory (not shown) that can be 100 to 1000 times larger than the on-chip memory—and induce a hierarchical bus architecture (not shown) to relay data to cores. GPU (Nvidia A100, 54 billion transistors, 7-nm process node, 60.25-megabyte on-chip memory) has 108 streaming multiprocessors (SMs) that share different partitions of the 40-megabyte L2 cache, with 192-kilobyte L1 cache or shared memory in each SM (*26*). TPU (Google TPUv4i, 16 billion transistors, 7-nm process node, 144-megabyte on-chip memory) has four cores with 128 megabytes of common memory (*27*). CPU (AMD Zen 3, 4.15 billion transistors, 7-nm process node, 36.5-megabyte on-chip memory) has a 32-megabyte L3 cache shared by all eight cores, with much smaller L2 and L1 caches in each core (*28*). (**B**) NorthPole has no centralized memory, instead distributing on-chip memory across the core array to colocate memory near compute (axiom 4). NorthPole (22 billion transistors, 12-nm process node, 224-megabyte on-chip memory) has a 16-by-16 array of cores (axiom 3) with enough on-chip memory (axiom 4) to store entire networks for many applications, without loading weights or instructions from off-chip memory during computation. The 16 framebuffer columns at the top of the NorthPole chip allow on-chip computation of the current frame to overlap off-chip communication of other frames. (**C**) Data flow diagrammed for accelerators such as CPU, GPU, and TPU, which have a hierarchy of off-chip memories, such as high-bandwidth memory (HBM) or dynamic random access memory (DRAM), and centralized on-chip caches. The leftmost column (Application) lists the sequence of operations, with the corresponding memory states and accesses in the columns to the right. Depending on the network size, either an off-chip memory located on the accelerator card or an on-chip cache stores instructions and weights (dashes) for the entire network. For each frame, the host application puts input tensor(s) of activations (boxes) into the off-chip memory to be conveyed to compute cores via on-chip caches. Memory misses during the computation of the network trigger cores to repeatedly fetch data from more-distant levels of the memory hierarchy. For each layer, the data move back and forth between cores and memory. (**D**) NorthPole stores the entire network on-chip, using prescheduling to ensure that there are no memory misses. For each frame, the host application writes input tensor(s) of activations (boxes) into the on-chip framebuffer to be conveyed to compute cores and, at the end of network computation, reads the output tensor(s). NorthPole handles scheduling of layers and data movement internally while maintaining neuron activations entirely within the core array. There is no layer-by-layer interaction between the host and NorthPole, which has only three commands (write input tensor, run network, read output tensor) and thus appears as an active memory chip, supporting operation that is autonomous from the host computer (axiom 10).

without compromising inference accuracy (supplementary texts S9 and S10).

Axiom 9: Codesigned software (fig. S3) automatically determines an explicit orchestration schedule for computation, memory, and communication to achieve high compute utilization in space and time while ensuring that there are no resource collisions. Network computation is broken down into layers, and each layer computation is spatially mapped to the core array. To minimize long-distance communication, spatial locality of neural activations is maintained across layers when possible. To prevent resource idling, core computation and intra- and intercore data movement are temporally synchronized so that data and programs are present on each core before use. Together, algorithms and software constitute an end-to-end toolchain that exploits the full capabilities of the architecture while providing a path to migrate existing applications and workflows to the architecture.

Axiom 10: NorthPole employs a usage model that consists of writing an input frame and reading an output frame (Figs. 1D and 3), which enables it to operate independently of the attached general processor (*16*). Once NorthPole is configured with network weights and an orchestration schedule, it is fully self-contained and computes all layers on-chip, requiring no off-chip data or weight movement. Thus, externally, the entire NorthPole chip appears as a form of active memory with only three commands (write input, run network, read output), with the minimum possible input-output bandwidth requirement; these features make NorthPole well suited for direct integration with high-bandwidth sensors, for embedding in computing and IT infrastructure, and for real-time, embedded control of complex systems (*22*).

Subsets of these axioms are found in previous architectures; however, as a whole, they are distinctive to the NorthPole architecture (Figs. 1, B and D; 2; and 3), which delivers outstanding performance on the fundamental metrics of energy, space, and time (Fig. 4) at state-of-the-art inference accuracy. The whole is greater than the sum of the parts. Details of the architecture, usage model, and translation of axioms into quantitative architecture design choices are in the supplementary text.

Silicon implementation

NorthPole has been fabricated in a 12-nm process and has 22 billion transistors in an 800-mm^2 area, 256 cores, 2048 (4096 and 8192) operations per core per cycle at 8-bit (at 4- and 2-bit,

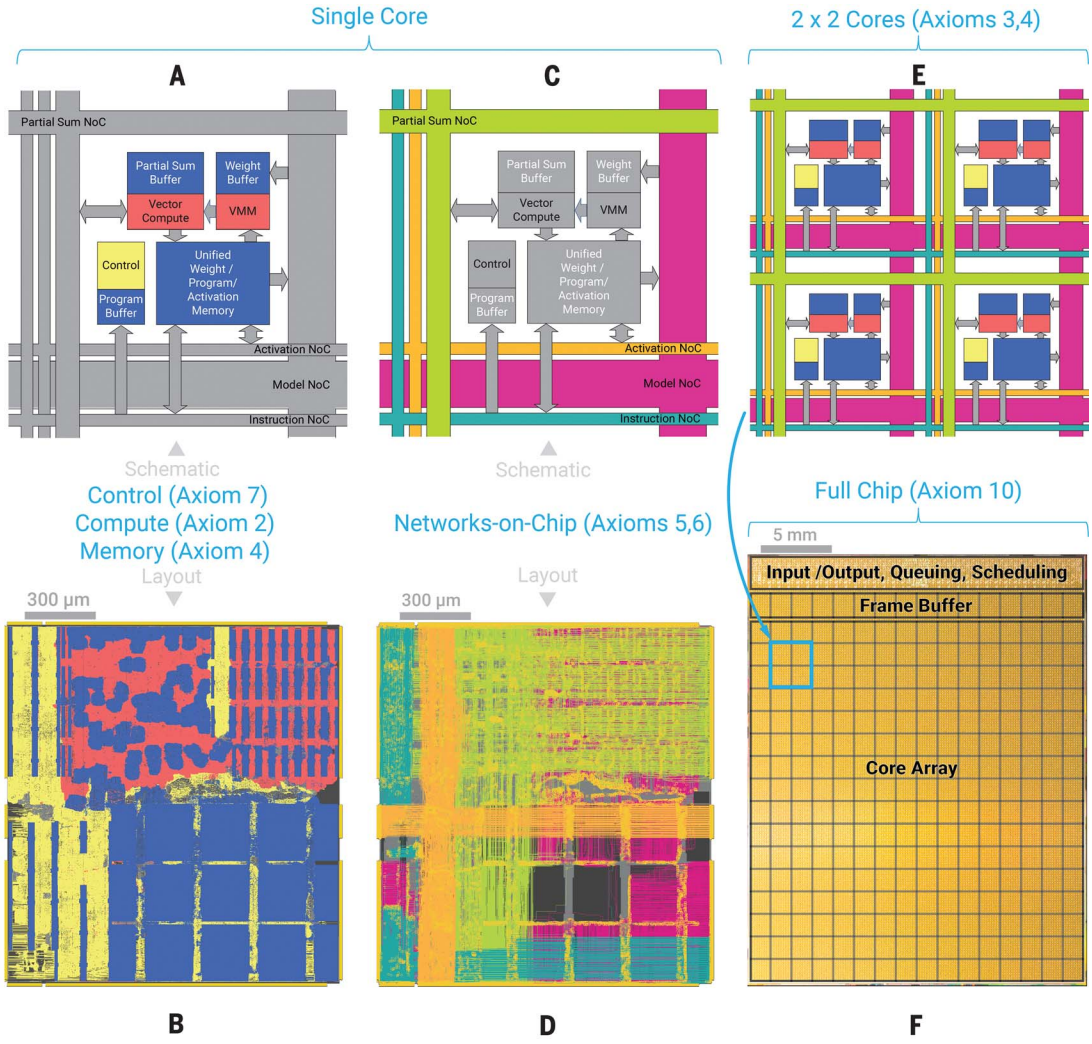


Fig. 2. NorthPole axiomatic architecture and implementation. (**A** and **B**) Core architecture schematic (A) and corresponding core circuit physical layout (B): Compute (red) includes the vector-matrix multiplier (VMM) and vector compute, which is integrated with memory (blue), and is driven by eight control threads (yellow). VMM supports 8-, 4-, and 2-bit precisions (axiom 2). Memories for weights, programs, and neuron activations are shared and unified. The unified memory is placed near the VMM and vector compute units (axiom 4). The physical placement of the VMM, vector units, and control units is intertwined, respectively, with weight buffer, partial sum buffer, and program buffer memories (axiom 4). At each cycle on each core, the VMM multiplies an input activation vector from the unified memory with a weight matrix from the weight buffer and transmits the output to the vector units, which in turn can communicate with the vector units on adjacent cores and, eventually, transmit the output to an activation function unit that, finally, writes the output activations back into the unified memory. The control unit has eight concurrent threads for orchestrating compute and communication (axiom 7). (**C** and **D**) Core NoC schematic (C) and core wire physical layout (D): A 512-wire partial sum NoC (green) transfers values between vector compute units in adjacent cores, enabling spatial computing (axiom 5); a 256-wire activation NoC (orange) transfers input frames to the cores from the framebuffer, transfers output frames from the cores to the framebuffer, and enables shuffling neuron activations among the unified memory of cores (axiom 5); a 1024-wire model NoC (magenta) carries distributed synaptic weights to the weight buffer (axiom 6); and a 256-wire instruction NoC (teal) delivers distributed instructions to the program buffer (axiom 6). (**E**) Schematic of spatially tiled two-by-two core array illustrating core-to-core communication using NoCs. (**F**) Die photograph of the manufactured, fully functional NorthPole chip with a 25-mm–by–31.8-mm footprint, illustrating periphery (input and output, queuing, and scheduling; axiom 10), framebuffer memory (for input and output tensor storage; axiom 10), and a 16-by-16 core array (axiom 3) where each core is 1.53-mm–by–1.65-mm and has 768 kilobytes of memory per core (axiom 4).

respectively) precision, 224 megabytes of on-chip memory (768 kilobytes per core, 32-megabyte framebuffer for input-output), more than 4096 wires crossing each core both horizontally and vertically, and 2048 threads. It is also fully operational in the first silicon implementation (Fig. 2F), is now deployed in a PCIe form factor printed circuit board (fig. S2), and has an end-to-end software toolchain (fig. S3 and supplementary texts S8 to S15). Implementation details are provided in supplementary texts S16 to S19. On the benchmark ResNet50 network, NorthPole can operate at frequencies ranging from 25 to 425 MHz with nearly linear power scaling (fig. S4), making it suitable for a myriad of applications from cloud to self-driving cars and achieving an unprecedented absolute energy metric of >1000 frames per joule (red point in fig. S4), with respect to energy consumed just by the NorthPole processor. Implementations of ResNet50 with 8-, 4-, mixed 4/2-, and 2-bit layers lead to progressively higher throughput and higher energy efficiency (fig. S5) and to progressively lower memory usage (supplementary text S11).

Energy, space, and time

For methodological rigor that ensures a fair and level comparison of various implementations, it is critical that all evaluation metrics be independent of the details of the implementations, which can vary arbitrarily across architectures at the discretion of the designers. The architecture-independent goodness metric adopted here is that all implementations must be measured at state-of-the-art inference accuracy. The architecture-independent cost metrics adopted here are now introduced.

Turning first to energy, different integrated-circuit (IC) implementations have different throughputs [in frames per second (FPS)] at different power consumptions (in watts). Therefore, FPS per watt (equals frames per joule) is a widely used energy metric for comparing

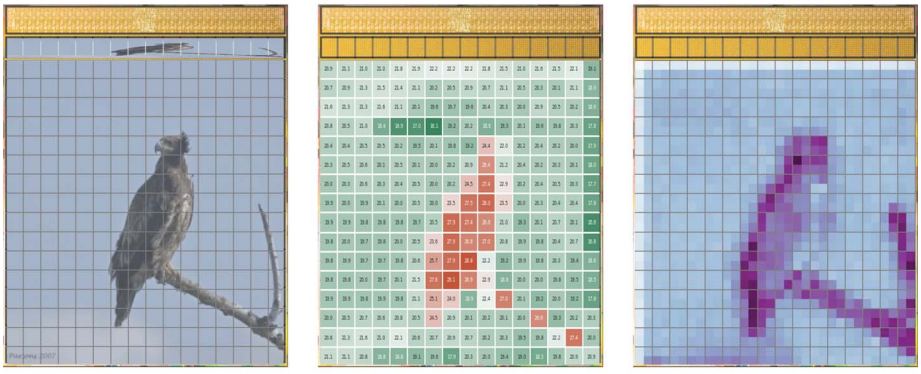


Fig. 3. NorthPole supports a spatial computing model for cortex-like topographic computing by exploiting network connectivity between adjacent cores. (**A**) An input image is divided into 16 subimages that are distributed to the framebuffer and then into 256 subimages that are distributed to the core array while retaining spatial structure (axiom 3). Intercore connectivity (axiom 5) enables receptive fields of neurons on a core to span nearby cores. [Photo credit: Raymond Parsons, used with permission] (**B**) The power consumption estimation in milliwatts by vector-matrix multipliers in the 16-by-16 NorthPole core array for the first convolutional layer of the ResNet50 network. Power consumption is computed by running a gate-level simulation of the NorthPole silicon design. The increased energy consumption shown in red matches the shape of the bird, demonstrating the in situ, activity-dependent neural computation of NorthPole that requires minimal communication between neighboring cores. (**C**) The average spatial intensity of tensor neuronal activation is shown for a layer about halfway through the layer-by-layer computation of the ResNet50 network. Comparing (C) with (A) demonstrates the slow spatial dispersion of neuron activity through the layers.

ICs. On this metric, Fig. 4A (*x* axis) compares prevalent architectures with NorthPole. Compared with the V100 GPU, which is implemented using a comparable 12-nm silicon technology process node and with a comparable number of transistors, NorthPole delivers 25 times more frames per joule. Many other compared architectures use silicon process nodes that are more advanced than 12 nm, and even though these architectures have an inherent advantage of using transistors with greater energy efficiency, NorthPole outperforms them. Even relative to the H100 GPU, which is implemented using a 4-nm silicon process node, NorthPole delivers five times more frames per joule. NorthPole has high energy efficiency because of less data movement, which is enabled by a lack of off-chip memory, the intertwining of on-chip memory with compute, the locality of spatial computing, and a better use of low-precision computation. Figure 4C decomposes FPS per watt into a product of FPS and 1/watt to illustrate through iso-energy-efficiency curves that, for prevalent architectures, higher FPS or lower power come, respectively, at the cost of higher power or lower FPS. Architecture trumps Moore's law.

Turning next to space, different ICs occupy different areas in different technology processes

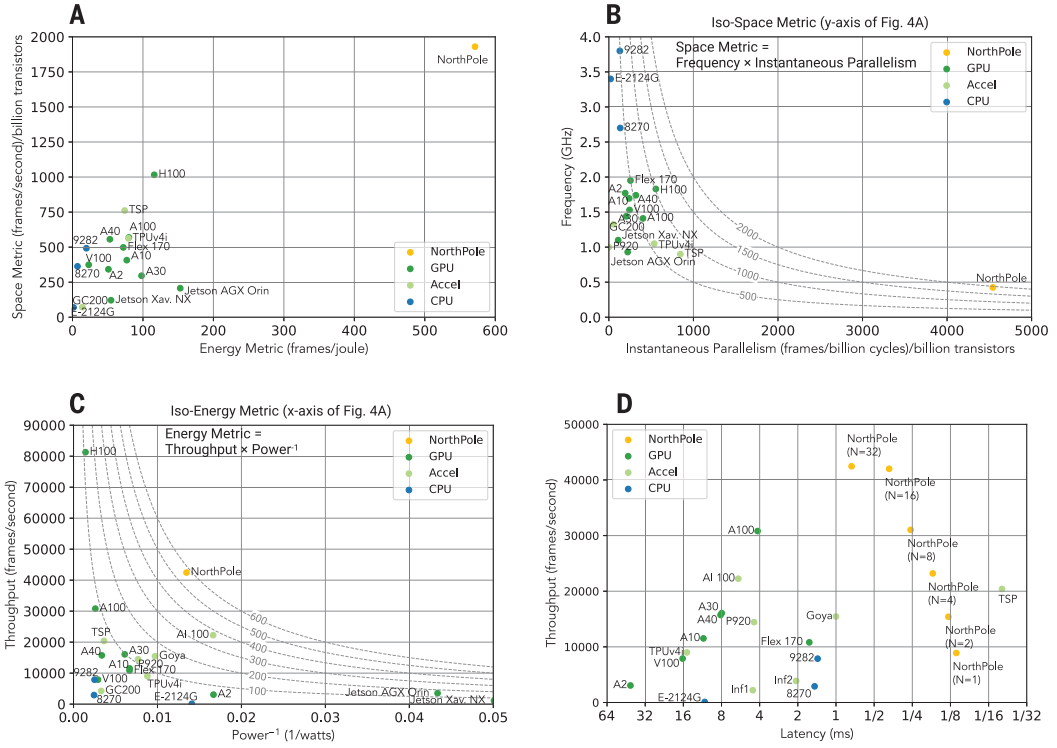


Fig. 4. NorthPole outperforms prevalent architectures on energy, space, and time metrics on ResNet50 at state-of-the-art inference accuracy. (**A**) NorthPole's space metric (FPS per billion transistors) and energy metric (frames per joule, equal to FPS per watt) exceed those of contemporary GPUs, CPUs, and accelerators (Accel). (**B**) Decomposition of FPS per billion transistors into the product of clock rate (cycles per second) on the *y* axis and normalized instantaneous parallelism [(frames per cycle) per billion transistors] on the *x* axis. The isoparametric curves denote frontiers of constant FPS per billion transistors. NorthPole performance derives from highly utilized parallelism, in contrast to both high–clock frequency CPUs (blue cluster) as well as parallelized GPUs (green cluster). (**C**) Decomposition of frames per joule into the product of FPS on the *y* axis and the inverse of power (1/watt) on the *x* axis. The isoparametric curves denote frontiers of constant frames per joule. (**D**) NorthPole throughput (FPS) on the *y* axis versus latency (ms) on the *x* axis in reversed log scale. NorthPole throughput and latency are reported at various batch sizes of 1, 2, 4, 8, 16, and 32. Compared with most comparative architectures, NorthPole achieves lower latency at similar throughput or, alternatively, higher throughput at lower latency. TSP (Tensor Streaming Processor) (*23*) is optimized for low latency but has substantially lower energy and space efficiencies. Power for all architectures includes total board power. For the details of each data point, see Table 1.

Table 1. ResNet50 performance comparison of neural inference across processors. Comparative processor performance obtained from published results that ran the ResNet50 network for image classification (*8*). Measured power includes total board power for all processors, and all networks are at state-of-the-art inference accuracy. The space and time-space metrics are calculated values. The energy metric is a reported value when available, otherwise it is calculated as FPS divided by thermal design power. The latency metric is a reported value when available, otherwise it is calculated as batch size divided by FPS. References and methodology for comparative performance numbers are provided in table S1. Accel, accelerator; CGRA, coarse-grain reconfigurable array; NIPU, neural inference processing unit; NA, not available.

	Design choices				Measured quantities		Computed figures of merit				Qualitative			
Processor	**Process node (nm)**	**Transistors (billions)**	**Frequency (GHz)**	**Batch size (frames)**	**Power (W)**	**Throughput (FPS)**	**Energy (frames per joule)**	**Space (FPS per billion transistors)**	**Time-space (frames per billion cycles per billion transistors)**	**Latency (ms)**	**Precision**	**Specialized to neural inference**	**Off-chip memory**	**Architecture category**
Xeon 9282	14	16	3.800	11	400	7878	20	492	130	1.4	INT8	No	Yes	CPU
Xeon 8270	14	8	2.700	NA	205	2906	7	363	134	1.5	INT8	No	Yes	CPU
E-2124G	14	1.4	3.400	NA	71	98	1.4	70	21	10.9	INT8	No	Yes	CPU
TPUv4i	7	16	1.050	16	175	9005	79	563	536	15.0	BF16	Yes	Yes	Accel:NIPU
Goya	16	NA	NA	10	103	15,453	149	NA	NA	1.0	INT8	Yes	Yes	Accel:NIPU
AI 100	7	NA	NA	8	60	22,250	370	NA	NA	5.9	INT8	Yes	Yes	Accel:NIPU
K200	14	NA	1.100	NA	160	NA	NA	NA	NA	NA	INT8	Yes	Yes	Accel:NIPU
Inf1	NA	NA	NA	10	NA	2207	NA	NA	NA	4.5	FP16	Yes	Yes	Accel:NIPU
Inf2	NA	NA	NA	8	NA	3878	NA	NA	NA	2.1	BF16	Yes	Yes	Accel:NIPU
TSP	14	27	0.900	1	275	20,400	74	761	846	0.049	INT8	No	No	Accel:CGRA
SN10	7	40	NA	NA	NA	NA	NA	NA	NA	NA	NA	No	Yes	Accel:CGRA
GC200	7	59	1.325	NA	300	4250	14	72	54	NA	FP16	No	Yes	Accel:Manycore
P920	12	NA	1.000	64	130	14,442	111	NA	NA	4.4	INT8	No	Yes	Accel:Manycore
WSE-2	7	2600	1.100	NA	15,000	NA	NA	NA	NA	NA	NA	No	No	Accel:Manycore
Flex 170	6	21.7	1.950	NA	150	10,811	72	498	255	1.6	INT8	No	Yes	GPU
V100	12	21.1	1.530	128	300	7896	23	374	244	16.2	INT8	No	Yes	GPU
Jetson Xavier NX	12	9	1.1	NA	20	1092	55	121	110	NA	INT8	No	Yes	GPU
Jetson AGX Orin	8	17	0.939	NA	40	3526	153	207	223	NA	INT8	No	Yes	GPU
T4	12	13.6	1.590	128	70	5003	72	368	231	25.6	INT8	No	Yes	GPU
A2	8	8.9	1.770	128	60	3059	51	342	193	41.8	INT8	No	Yes	GPU
A10	8	28.3	1.695	128	150	11,520	77	407	240	11.1	INT8	No	Yes	GPU
A40	8	28.3	1.740	128	300	15,727	53	556	320	8.1	INT8	No	Yes	GPU
A30	7	54.2	1.440	128	165	16,057	98	296	206	8.0	INT8	No	Yes	GPU
A100	7	54.2	1.410	128	400	30,814	80	569	404	4.2	INT8	No	Yes	GPU
H100	4	80	1.830	NA	700	81,292	116	1016	555	NA	INT8	No	Yes	GPU
NorthPole	12	22	0.425	32	74	42,460	571	1930	4541	0.75	INT8, 4, or 2	Yes	No	Accel:NIPU

but can be compared in terms of the number of transistors. Therefore, FPS per billion transistors is a meaningful space metric for comparing ICs. On this metric, Fig. 4A (y axis) demonstrates that NorthPole outperforms prevalent architectures and that NorthPole delivers five times more FPS per billion transistors than the V100 GPU. It is notable that an architecture that devotes substantially more transistors to memory than logic outperforms prevalent architectures that devote substantially more transistors to logic than memory, on a performance-per-transistor space metric. To throw light on this advantage, observe that the space metric, FPS per billion transistors, can be decomposed into a product of normalized instantaneous parallelism [(frames per cycle) per billion transistors] and clock rate (cycles per second). All compared architectures implement considerably faster clock rates than NorthPole, and yet NorthPole outperforms on the space metric by achieving substantially higher instantaneous parallelism (through high utilization of many highly parallel compute units specialized for neural inference) and substantially lower transistor count owing to low precision (Fig. 4B).

Turning finally to time, different ICs permit different numbers of input frames to be bundled for processing in a batch, so batch size per FPS, or latency, is an appropriate time(-to-solution) metric. NorthPole throughput and latency are reported at various batch sizes of 1, 2, 4, 8, 16, and 32, thus providing an entire throughput-latency frontier that envelops most of the prevalent architectures (Fig. 4D). For example, with a batch size of 32, NorthPole can achieve a high throughput of 42,460 FPS at a latency of 753 µs, and, with a batch size of one (a single image), NorthPole can operate at a low latency of 106 µs at a throughput of 9454 FPS [(*23*) reports a processor with lower latency but also lower energy and space efficiencies]. NorthPole achieves a lower latency because it has a higher FPS owing to massive parallelism, has low latency local memory access, and hides communication latency behind compute operations, and because it can compute with much smaller batch sizes while maintaining high utilization owing to spatial computing.

The results shown in Figure 4 derive from the ResNet50 image classification network, which is widely used as a benchmark. Similar comparative results can be seen for the Yolo-v4 detection network (fig. S6).

Applications

NorthPole's on-chip memory is sufficient to implement many widely used networks across a variety of application domains; a few canonical examples are the ResNet family (for classification), Yolo and SSD-VGG (for detection), PSPNet (for segmentation), RetinaMask (for instance segmentation), BERT (for natural language processing), and DeepSpeech2 (for speech recognition). For examples of networks that are implementable on a single NorthPole chip, see supplementary text S12 and tables S7 to S12. With 4- or 2-bit precisions, the number of neural network weights stored on NorthPole can be doubled or quadrupled, respectively. Memory permitting, more than one network can be resident in the on-chip memory, thus supporting multitenancy. Inspired by the brain, which has no off-chip memory, NorthPole is optimized for on-chip networks, but, if so desired, larger networks can be broken down into smaller subnetworks that do fit in NorthPole's model memory, and these subnetworks can be pipelined on multiple NorthPole chips (supplementary text S25). Given the present silicon technology roadmap (*10*), 50 times more memory on-chip seems possible by going from 37 million transistors per square millimeter on the present 12-nm implementation to 2 billion, thus enabling much larger networks to fit on-chip.

Conclusions

As seen from the inside of the chip, at the level of individual cores, NorthPole appears as memory near compute; as seen from the outside of the chip, at the level of input-output, it appears as an active memory. NorthPole is an architectural innovation at the intersection of brain-inspired computing and semiconductor technology, which defines a frontier that promises to expand. Specifically, the NorthPole trajectory promises to evolve along dimensions of algorithms (natively optimized for NorthPole), systems (scale-out with multiple boards, scale-up with multiple chips per board, and direct sensor integration), modularity (chiplets with varying core array sizes from tiny to large), packaging (heterogeneous chiplet integration), architecture (tileability, three-dimensional integration), silicon scaling (*10*), silicon optimization (*24*), and, potentially, post-silicon technologies (*25*). Interestingly, the EDVAC report began with neurons and synapses, and now the rising importance of neural inference applications has brought the field full circle.

REFERENCES AND NOTES

1. J. von Neumann, *IEEE Ann. Hist. Comput.* **15**, 27–75 (1993).
2. O. Mutlu, S. Ghose, J. Gómez-Luna, R. Ausavarungnirun, in *Emerging Computing: From Devices to Systems*, Computer Architecture and Design Methodologies series, M. M. S. Aly, A. Chattopadhyay, Eds. (Springer, 2023).
3. W. A. Wulf, S. A. McKee, *Comput. Archit. News* **23**, 20–24 (1995).
4. M. Horowitz, in *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* (IEEE, 2014), pp. 10–14.
5. J. J. Dongarra, *Commun. ACM* **65**, 66–72 (2022).
6. J. Backus, *Commun. ACM* **21**, 613–641 (1978).
7. T. J. Sejnowski, P. S. Churchland, *The Computational Brain* (MIT Press, 1994).
8. J. Deng *et al*., in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009), pp. 248–255.
9. T. J. Sejnowski, *The Deep Learning Revolution* (MIT Press, 2018).
10. S. Datta, W. Chakraborty, M. Radosavljevic, *Science* **378**, 733–740 (2022).
11. R. H. Norman, "Solid state switching and memory apparatus," US Patent 3,562,721 (1971).
12. S. S. Iyer *et al*., *IBM J. Res. Develop.* **49**, 333–350 (2005).
13. N. C. Thompson, K. Greenewald, K. Lee, G. F. Manso, arXiv: 2007.05558 [cs.LG] (2020).
14. P. A. Merolla *et al*., *Science* **345**, 668–673 (2014).
15. C. Mead, M. Ismail, Eds., *Analog VLSI Implementation of Neural Systems* (Springer, 1989).
16. C. E. Leiserson *et al*., *Science* **368**, eaam9744 (2020).
17. T. M. Bartol Jr. *et al*., *eLife* **4**, e10778 (2015).
18. V. B. Mountcastle, *Brain* **120**, 701–722 (1997).
19. S. S. Iyer, *MRS Bull.* **40**, 225–232 (2015).
20. T. Bjerregaard, S. Mahadevan, *ACM Comput. Surv.* **38**, 1 (2006).
21. M. I. Sereno, M. R. Sood, R.-S. Huang, *Front. Syst. Neurosci.* **16**, 787737 (2022).
22. J. Degrave *et al*., *Nature* **602**, 414–419 (2022).
23. D. Abts *et al*., in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)* (IEEE, 2020), pp. 145–158.
24. F. Khan, E. Cartier, J. C. S. Woo, S. S. Iyer, *IEEE Electron Device Lett.* **38**, 44–47 (2017).
25. W. Wan *et al*., *Nature* **608**, 504–512 (2022).
26. J. Choquette, E. Lee, R. Krashinsky, V. Balan, B. Khailany, in *IEEE International Solid-State Circuits Conference (ISSCC)* (IEEE, 2021), pp. 48–50.
27. N. P. Jouppi *et al*., in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)* (IEEE, 2021), pp. 1–14.
28. M. Evers, L. Barnes, M. Clark, *IEEE Micro* **42**, 7–12 (2022).

ACKNOWLEDGMENTS

We thank many collaborators: I. Ahmed, T. Amberg, N. Antoine, F. Baez, H. Baier, L. Berge, R. Bhimarao, J. Bjorgaard, M. Boraas, B. Brezzo, M. Butterbaugh, E. J. Campbell, E. Colgan, M. Criscolo, C. di Nolfo, M. Doyle, M. Grassi, C. Guido, D. Hitchings, R. Hoke, K. Holland, S. Hui, T. Jones, S. Kedambadi, P. Kourdis, A. LaPiana, G. LaPiana, S. Lekuch, Z. Liu, R. Loboprabhu, A. Lonkar, D. Ma, G. Maass, S. Machha, V. Mahadevan, P. Mann, M. Mastro, K. O'Connell, B. Parikh, H. Penner, R. Purdy, M. Rajwadkar, L. Rapp, K. Rice, C. Sassano, K. Schoneck, R. Scouller, J. Shaffer, D. Shukla, A. Sigler, C. Smallwood, N. Subbiah, S. Tian, T. Tidwell, D. Turnbull, M. Uman, G. Van Leeuwen, S. Vispute, I. Vo, J. Wang, C. Werner, S. Wundavilli, and C. N. Xiong. We also thank H. Chandran, J. Saxena, J. Lee, A. Prashantha, S. Rao, and the Anora Labs DFT team as well as teams from Cadence Design Systems and Globalfoundries. We thank three anonymous reviewers for their thoughtful comments. **Funding:** This material is based upon work supported by the United States Air Force under contract no. FA8750-19-C-1518. **Author contributions:** Conceptualization: D.S.M.; Architecture: R.A., J.V.A., A.S.C., D.S.M., and J.Sa. (co-leads), F.A., A.Am., A.An., P.D., M.V.D., S.K.E., N.J.M., T.K.N., C.O.O., B.T.; Algorithms: S.K.E. (lead), R.A., D.B., J.L.M., D.S.M.; Compiler: P.D. (lead), A.Am., R.A., D.B., S.K.E., R.G., N.J.M., J.L.M., D.S.M., T.K.N., B.T., I.T.; Validator: A.An. (lead), N.J.M., D.S.M., T.K.N., B.T.; Runtime: J.A.K. (lead), A.An., R.A., A.S.C., M.V.D, R.G., D.S.M., D.F.S., B.T.; Logic design: A.S.C. (lead), F.A., G.J.G., M.I., C.O.O., J.Sa., J.Si., T.U.; Logic verification: M.V.D., R.A., and D.S.M. (co-leads), A.An., A.Am., D.B., P.J.C., P.D., S.K.E., M.D.F., R.G., G.J.G., M.I., J.A.K., N.J.M., J.L.M., T.K.N., W.P.R., B.S., J.Si., J.Sa., D.F.S., B.T., I.T., T.U.; Physical design: F.A. and C.O.O. (co-leads), J.V.A., A.S.C., Y.N., J.Sa., T.U.; Design for test: J.Sa. (lead), J.V.A., A.S.C., G.J.G., M.I., C.O.O., J.Si., T.U.; Board design: F.A. and K.S. (co-leads), J.V.A., A.S.C., M.V.D., C.O.O., W.P.R.; Test and measurement: R.A., J.V.A., A.S.C., M.V.D., C.O.O., and J.Sa. (co-leads), F.A., A.Am., A.An., D.B., P.J.C., P.D., S.K.E., M.D.F., R.G., G.J.G., J.A.K., M.I., N.J.M., J.L.M., D.S.M., T.K.N., W.P.R., B.S., J.Si., K.S., D.F.S., B.T., I.T., T.U.; Resources: M.V.D., M.D.F., G.J.G., J.A.K., J.L.K., N.J.M., D.S.M., W.P.R.; Supervision: D.S.M. served as the principal investigator. J.V.A. and J.Sa. supervised hardware. R.A., M.D.F., and B.T. supervised software. K.S. supervised board design; Project administration: J.L.K. (lead), P.J.C., D.S.M., W.P.R., B.S.; Funding acquisition: D.S.M. (lead), J.L.K. **Competing interests:** The authors are researchers at IBM, which has a commercial interest in the NorthPole intellectual property and technology. **Data and materials availability:** All data needed to evaluate the conclusions in this study are available in the main text or the supplementary materials.

License information: Copyright ©

2023 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adh1174
Supplementary Text
Figs. S1 to S8
Tables S1 to S12
References (*29*–*104*)

Submitted 17 February 2023; accepted 1 September 2023
10.1126/science.adh1174

WORKING LIFE

By **Roel Snieder**

Just be kind

I was a junior professor, stressed and working myself to exhaustion yet still feeling inadequate both at work and at home. So, when three students from my introductory programming course came to my office at the end of the semester with concerns about their grades, I didn't feel I had time to deal with their complaints. I felt they simply hadn't learned the material, and I told them to leave my office. Months of unpleasantness followed as the issue escalated and eventually came before a university arbitration committee. In hindsight I could see I should have listened to the students' concerns. But I was too wrapped up in judgment, competition, and overwork to be kind.

As I pursued my research career, I operated from the premise that to contribute to science I had to give it everything—and I expected the same from those around me. Even though I had been trained by advisers who combined brilliance with care, I had bought into the mindset that only the fittest survive in science. It had brought me success: a solid track record of publications and funding, and a faculty position. I was also chronically tired and plagued by feelings of insufficiency when I compared myself with colleagues who were more productive and other parents who seemed to be better with their kids—but I didn't see another way to proceed.

"I was too wrapped up in judgment, competition, and overwork."

The turning point came a few years later during a conversation with a good friend who works outside academia. When I told them about my struggles with overwork and inadequacy, they asked, "Does the fact that you can do better mean that you are not doing a good job?" That prompt helped me see that, apart from the occasional misstep, I was doing well at my job and that my children were growing up in a loving home. I realized that what needed to change was not my work, but my mindset. I had to learn to accept that perfection is an illusion and that there is a place for compromise.

This shift made me reflect on how I wanted to behave as a scientist and a teacher. Was this career about racking up the largest number of publications? Did I want to focus my teaching on the brightest students while ignoring the needs of others? The answer to both questions was negative; yet my actions said otherwise. I needed to be more intentional about making my career meet my personal priorities.

This launched me on a decadelong effort to shift my behavior. One strategy that helped was practicing mental hygiene, an approach I learned through a psychiatrist friend and classes at an interfaith church. It taught me to monitor my thinking, eliminate toxic thoughts, and choose what mindset I want to embrace. I began to make a point of empowering others rather than judging them, finding mentors and role models to guide me and hold me accountable, refraining from gossip, and building meaningful relations as a foundation of my academic career. It wasn't always easy, but changing my mindset was key in counteracting the negativity that can permeate academic culture, and that I had taken on.

I've also developed three concrete work habits that enable me to realize my intentions while also furthering my research and career: building personal connections, sharing ideas freely, and being a reliable and fun collaborator. I've seen how working with colleagues with whom I have a personal connection unleashes creativity and progress. Such connections don't form overnight. But when I take the time to nurture them by sharing meals and personal conversations—which earlier in my career I might have rushed through or not done at all—I know that is time well spent. And my productivity hasn't suffered; if anything, my research has benefited from the increased creativity and engagement.

Sometimes I fail to live up to my intentions for the simple reason that I forget and get carried away. Recently, for example, I found myself snapping at a colleague when I felt overwhelmed by bureaucratic demands in the busiest time of the semester. The good news, however, is that we don't need to be perfect to be good. What I've learned in my journey of decades can be summarized in a single sentence: Just be kind. ■

Roel Snieder is a professor at the Colorado School of Mines.
Send your career story to SciCareerEditor@aaas.org.